Statistical models for short-term animal behaviour

David John Allcroft

Doctor of Philosophy
University of Edinburgh

2001

Abstract

This thesis aims to identify appropriate methods for the modelling of short-term animal behaviour data and, in the wider context, any time series of categorical data. Extensive use is made of a large dataset of cow feeding behaviour, consisting of full feeding records for a number of cows over a relatively long period of time, the data taking the form of binary time series, i.e. feeding/non-feeding periods. After initial exploratory data analysis, I go on to investigate three main classes of model — latent Gaussian, hidden Markov and semi-Markov.

The latent Gaussian model assumes the binary data occur from the thresholding of an underlying continuous variable. I identify the one-to-one relationship between the autocorrelation of the observed and latent variables and consider techniques for parameter estimation. For a multivariate stationary Gaussian process, I show the asymptotic equivalence of the likelihood in its spectral and conventional forms, and provide a proof that for short-term memory processes such as ARMA models, a good approximation for the spectral form can be obtained using Fourier transforms of correlations at only the first few lags. A simulation study highlights the saving in computing time that this offers, and also shows that, in contrast to the least squares methods considered, the number of lags to retain is not crucial for obtaining efficient parameter estimates.

The attractiveness of hidden Markov models for behaviour data is also due to the direct modelling of the underlying state of the animal, but the latent variable here is discrete, following a Markov chain. Observations are conditionally independent of each other, dependent only on the current state of the Markov chain. However this type of model constrains the durations between feeding events to follow a mixture of geometric distributions. This is seen to be inappropriate for the cow feeding data, mixtures of log-normal distributions offering a better description both statistically and biologically.

Semi-Markov models involve the animal moving between a set of feeding and non-feeding states according to a set of transition probabilities, the marginal distributions for durations in each state being specified directly. The semi-Markov models fit here have more than one non-feeding state and so the current state of the animal when not feeding is unobserved. I therefore generalise the basic estimation procedure for semi-Markov models, using a form of the EM algorithm to allow this uncertainty to be taken into account.

I compare the models overall in terms of their existence in discrete and continuous time, types of latent structure assumed, marginal distributions of feeding and non-feeding durations and time-dependence. Formal model comparisons need to take account of the models being non-nested and fit according to different criteria, and so a parametric bootstrap approach is developed, involving simulation from each of the fitted models and the subsequent re-fitting of all models to each simulated dataset. Comparison of the fitting criteria for the observed and simulated datasets can then be used to decide how likely it is for the data to have arisen from each model. Bearing in mind that generalisations cannot be made for all types of behavioural data, it is concluded that, of the models investigated, the semi-Markov offers the most appropriate description for the cow feeding data.

Acknowledgements

I would like to thank Chris Glasbey, my main supervisor, for all his help and guidance throughout this project. I also thank my other supervisors, Ilias Kyriazakis, Colin Aitken and Elizabeth Austin for their help and, in addition, Bert Tolkamp for many useful discussions. Thanks also to all the staff and students at BioSS, the Scottish Agricultural College and the Mathematics and Statistics Department of Edinburgh University. I gratefully acknowledge BioSS/ICMS for my research studentship and Langhill Dairy Cattle Research Centre for providing the data. Finally I would like to thank Stephen Baillie and Jayne McIntyre for help with proof-reading, but more importantly I am grateful to them and others for much encouragement and support throughout.

Table of Contents

Chapter 1

Introduction

In this introductory chapter, the motivation for the project as a whole is discussed and I introduce the data, the literature and the models that form the starting point for the work in later chapters. After setting out the main objectives in Section 1.1, Section 1.2 introduces the cow feeding dataset that is considered throughout and discusses what particular challenges are posed in modelling these type of data. Section 1.3 presents a short review of the animal behaviour literature, outlining some of the concepts that have previously been applied in this area. Next, Section 1.4 reviews some statistical models which form the basis for the models I develop later, and finally the layout of the remainder of the thesis is summarised in Section 1.5.

1.1 Motivation and objectives

There is a considerable amount of work in the literature on the long-term behaviour of animals, but relatively little on short-term behaviour. Work on long-term behaviour might involve looking at the overall behaviour profiles of animals, or at whether goals have been achieved over a specified period. Sufficiently long periods of time need to be considered and interest lies mainly in summaries of the data over these long periods. For the study of short-term behaviour, emphasis is instead on how the individual behavioural events occur. I want to try and improve understanding of how an animal makes decisions about what behaviours to perform, and hope to do this by identifying useful modelling approaches that reflect the animal's underlying motivation for performing given behaviours.

In terms of feeding, over a long period it is apparent that in order to survive and remain in good condition, animals must achieve certain goals in terms of amount

and type of intake. But how are feeding events organised in the short term? As long as a cow satisfies her weekly nutrition requirements, is she bothered on a day to day basis about getting a particular amount of food? Within a day is she bothered about how many separate meals she has and their size, composition and distribution over the day? It might be the case that as long as she satisfies her long-term nutrition requirements, the details of how she does this are of no consequence, and so it does not matter to her whether she has six or four meals in a day. Alternatively she might be very specific about her meal patterns.

Every behaviour an animal performs can be thought of as being the result of various internal and external pressures, such as environmental and physiological factors. Therefore motivation to feed can simply be thought of as a function of current and previous conditions and events. Kyriazakis (1997) considered the feeding behaviour of farm animals in terms of achieving goals, concluding that long-term feeding behaviour is closely related to the long-term changes in the animal's internal state, whereas short term behaviour may simply be organised to make efficient use of the immediate environment and be dictated by habit. It is then only when a large change in the animal's internal state occurs, as opposed to the usual short-term fluctuations, that the animal is prompted to change its behaviour. Hence if an animal is going into a state of deficit with regards to feeding, there will come a critical point when she has to eat.

A related idea, which is discussed more fully later on, is that of *latent states*, and all the models we develop in later chapters have a latent component to them. Observed behaviours can change at a relatively fast rate, but it is not necessarily important to focus on this, as it may not be a good reflection of the overall current state of the animal. I therefore conjecture the existence of an underlying slower-changing series of states that the animal is moving through. Consider words such as 'feeding', 'resting' and 'active'. These sorts of words may represent 'states of mind' of the animal, rather than observable behaviours. Within this sort of framework, an animal could remain in an overall resting state, but exhibit small amounts of behaviours that are more usually associated with other states. This idea is an important one, as a biological model might be more useful if it models these underlying states rather than every small observed change in behaviour.

The main objectives of this project are both biological and statistical.

- The main biological objective is to improve understanding of how animals organise their behaviour in the short term, and relate this to the fulfillment of longer-term goals. By identifying suitable models and methods of infer-

ence, biologists can formulate plausible biological mechanisms that could have given rise to the observed data.

- As well as being of general scientific interest, animal welfare is an impetus for this work. If the behavioural patterns of content animals, housed in good conditions and on a quality diet, are known, this can be used as a benchmark for comparing animals housed in inferior conditions or fed a poorer diet, giving important information on whether a particular housing scheme or feeding regime is having a detrimental effect on the animals' behaviour and hence well-being. Behavioural datasets are often very large, so by fitting an appropriate model we can reduce a large amount of data to a relatively small set of parameters which provide a succinct summary of the data, and so enabling the easy comparison of different groups of animals.

- The cow feeding data take the form of binary time series, but the models we look at have the scope for generalising to more than two categories of behaviour. Therefore from a statistical viewpoint the project can be considered as an investigation into models applicable to time series of categorical data. So although the project is specific to animal behaviour in that any model developed should be biologically relevant, many of the techniques developed will be applicable to other types of data that have a similar structure. Hence statistically, I am interested in the modelling of general time series of categorical data and establishing sound statistical techniques for the estimation of parameters in such models. In addition, the comparison of non-nested models is an important issue and techniques will be developed for this.

1.2 Data

My work is centred around an extensive set of cow feeding data from the Langhill Dairy Cattle Research Centre, Roslin, Midlothian, taking the form of complete records of feeding behaviour for each of thirty-four cows over a 30 day period in April – May 1995. The cows had continuous access to a 70:30 mix of silage/ concentrate feed in computerised feeders. Transponders worn around the cows' necks enabled the feeders to automatically record information every time feeding occurred, including the time the cow accessed the feeder, the time she subsequently left the feeder and the weight of food left in the feeder on exit. Therefore we know the length of each feeder-visit, recorded to the nearest second, and the amount of food consumed during that visit. Detailed experimental details are

Figure 1.1: *Cows feeding from the automatic feeders at Langhill.*

given in Tolkamp and Kyriazakis (1997) and Tolkamp et al. (1998b). Figure 1.1 shows a picture of cows at the feeders.

Of the thirty-four cows, eight had access only to feeders containing a high-protein feed, ten just to feeders containing a low-protein feed and a third group of sixteen had access to both. This last group are termed *choice cows* and for this group records also include the type of food for that visit. Cow identification numbers will be used later on and so are given here:

- High protein (HP) – 5, 41, 108, 169, 170, 182, 194, 221,
- Low protein (LP) – 1, 9, 48, 66, 75, 77, 110, 118, 176, 224,
- Choice (CH) – 3, 37, 43, 47, 70, 76, 122, 132, 134, 150, 165, 171, 179, 197, 223, 237.

The data described are only part of a much larger dataset covering a longer period of time. The particular 30 day period was selected as one for which the number of animals in the yard was reasonably stable and the animals had already been there for some time and so were used to the environment and feeding regime. It was decided that this dataset provided adequate potential for the investigation and comparison of different modelling techniques and that further data would not be considered at this stage.

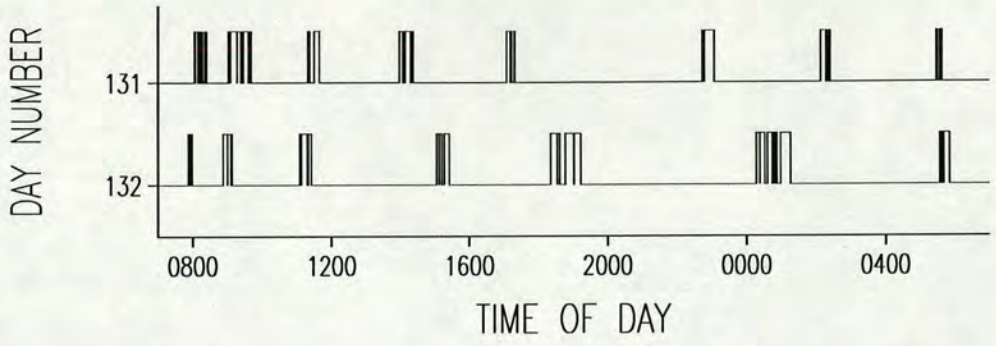

Figure 1.2: *Two days of feeder-visit data for Cow 41. Raised values of the signal denote periods of feeding.*

The data we model are in the form of binary time series, taking the value 0 when not feeding and 1 when feeding occurs. As an example, Figure 1.2 shows two days of such data for Cow 41, one of the animals on the high-protein diet. The cows on this diet are used as the main examples throughout this thesis, as being on a single, high-quality diet, they are likely to be the most straightforward group for which to develop models. Comparison with animals on the other regimes would be the next step after this. We present data from all eight of the high-protein cows in Appendix A, along with a small amount of data from animals on the other feeding regimes. The thirty days are numbered from 106 to 135, day 1 being 1 January 1995, hence the data cover the period 16 April to 15 May 1995.

One of the most noticeable features of the data is that feeding usually occurs in meals or *bouts*, each made up of several shorter individual visits to feeders. Inspection of the whole set of plots in Appendix A reveals that there are many differences between individuals. For example Cow 221 (Figure A.8) has many meals containing upwards of 10 individual visits, whereas for Cow 9 (Figure A.9), many meals are composed of only a single visit. We do not know how much these characteristics are a result of the individual cow's preference, and how much is imposed by the rest of the herd. For instance Cow 221 may choose to have many visits per meal, preferring to change feeders or take short breaks, or it may be the case that she is being bullied out of feeders by other cows. This motivates the theory that visit data are highly affected by dominance effects within the herd, and that it might be more appropriate to group individual visits into meals and model the resulting data, with the hope that this would allow better comparison with other herds in different conditions. The problem of identifying bouts is discussed in detail in Chapter 2. However, as the main interest here is in short-

term feeding behaviour, it would be useful for a model to be able to describe intra-meal as well as inter-meal behaviour. Therefore I choose to develop models for the visit data and, where appropriate, indicate how they can be adapted to model meal data instead.

Diurnal pattern also varies considerably between individuals. Cows typically have several bouts of feeding behaviour during the day and somewhat less feeding during the night. Typically around two-thirds of all intake is during the day (taken as the twelve hours from 08:00 – 20:00), with one third at night. However inspection of the data shows that there are clearly many differences between animals. For example, Cow 108 shows quite a strong diurnal pattern, generally having around five meals between the hours of 08:00 and 20:00, with one more meal during the night and another between 06:00 and 07:00. In contrast it is difficult to give any typical meal times for Cow 5. She appears to have very little diurnal pattern, and frequency of feeding during the night looks similar to that during the day. Again we don't know whether Cow 5 is choosing not to follow a regular daily pattern, or whether she is a less dominant cow and taking the opportunity to feed whenever she can or when particular other cows are not around to bully her. When considering the different models, inclusion of diurnal trends will be investigated.

The differences already noted between individuals makes the existence of a universal feeding strategy for all cows seem unlikely. This in turn questions whether any single model would be capable of describing the feeding patterns of all the cows here. Particularly it makes it unlikely that a multivariate approach, modelling all cows simultaneously, would be successful. The advantage of such a model would be that structure could be imposed which would allow common features across the dataset to be estimated using information from all animals simultaneously. However with the apparent differences between animals already noted, cows will be modelled individually in the hope of allowing enough flexibility to capture the features of each animal. In addition, the scope of the models for being extended to modelling groups of cows simultaneously will also be considered.

1.3 Review of animal behaviour literature

The short review presented in this section takes a broad look at areas of modelling in the animal behaviour literature, in order to identify suitable areas of work to progress. I am particularly interested in models applied to feeding behaviour, but applications to other behaviours may be equally relevant and similarly I

will not restrict the review to behaviour of cows. The models are discussed in terms of whether they are appropriate, valid, useful, interesting, etc., both from a biological and a statistical viewpoint. Comment is restricted mainly to papers that do not fall naturally into the areas covered by later work, in which case discussion is reserved for the relevant chapter. Finally it should be noted that I have concentrated on only a small number of papers that I have found interesting, there being many more similar papers which are not included.

1.3.1 General

As already noted for the cow feeding data, an important feature of animal behaviour is that it often occurs in bouts, i.e. several distinct occurrences of a particular behaviour close together in time, followed by large gaps in between. Methodology for deciding which gaps between behaviours are short enough to be considered within-bout and which should be taken as separating bouts has been developed over a number of papers, discussed in Chapter 2, but this idea is also part of some of the more general papers considered here.

Slater (1974) looked at the behaviour of zebra finches, presenting a thorough investigation into the empirical distributions of bout lengths and gap lengths, finding some evidence of correlation. The presence of marked diurnal patterns led to analysis being restricted just to the part of the day which displayed little diurnal variation. Individuals were treated separately, this being a sensible approach when there are substantial differences between individuals, otherwise it is possible to end up with an 'average' animal which is not typical of any. Hence as already discussed, it is often best to consider animals separately and then draw general conclusions at the end. Slater and Ollason (1972) also did this, fitting Markov models to see which behaviours are associated with each other via their transition probabilities. They argued that it is important that an approach retains the temporal nature of the data, however in their desire to keep recording of behaviour as detailed as possible, and so using fourteen categories of behaviour, investigation of models beyond a first order Markov process was not possible, as too little data were generally available. Simpson (1990) gave a general discussion on the feeding patterns of locusts, and addressed many of the issues which will be discussed later, including the distribution of inter-feed times, the subsequent definition of a meal and the description of the probability of the next meal starting in terms of the hazard function of the distribution of inter-meal durations. He also looked for relationships between meal size, duration and ingestion rate, and although the ideas presented are easily transferable to other animals, the findings here

for locusts may well not be; due to the obvious physiological differences, cows are quite likely to operate according to different mechanisms. Haccou and Meelis (1994) reviewed a wide range of techniques for animal behaviour based on Markov models, ranging from simple tests of homogeneity and exponentiality to the fitting of continuous-time Markov and semi-Markov processes. These are useful models for animal behaviour in that the time-structure of the data is a central feature. I describe these processes later in this chapter and look at the fitting of semi-Markov processes in detail in Chapter 6.

1.3.2 Automata and rule-based methods

Cellular automata (CA) models can be seen as simple models that approximate physical laws by a set of simple rules. Typically a cellular automaton is considered on a discrete grid, in discrete time, and the state of a particular cell at the next time point is determined by some rule based on the current states of adjacent cells. The most well-known example is the *Game of Life*, see for example Ermentrout and Edelstein-Keshet (1993), who reviewed CA models, emphasising that they are not a replacement for traditional mathematical models, but may be a helpful first step in the modelling process.

Thuijsman et al. (1995) looked at these ideas applied to foraging behaviour of bees. For two different colours of flowers, they discuss ideas such as the *ideal free distribution* (IFD), which is the expected distribution of the bees on the two colours given a large number of bees. The expected proportions are p and q, the probabilities for the two colours of obtaining nectar above the minimum acceptable level, i.e. the *critical level*. The *matching law* is a similar idea for the time allocated to each of the colours by a single forager. It is thought that bees use *bounded recall*, i.e. they only remember success/failure information from their last few flowers, rather than everything about their previous foraging, to make decisions about where to go next. Two sampling strategies are considered — e-sampling, where the bee stays with the colour it is currently on with probability $1 - e$ and moves to the other colour with probability e, the other strategy being for the bee to stay at its current colour until a certain number of failures, after which it will move to the other colour. Both strategies can result in the IFD. There are many more papers in this area, e.g. Harley (1981) and Houston and Sumida (1987), the theories mainly being applied to foraging or vigilance behaviours. Evolutionarily stable strategies (ESS) and developmentally stable strategies (DSS) were described by McNamara and Houston (1985), who looked at the conflicts between optimal foraging and learning and asked whether these

can both take place simultaneously, coming to the conclusion that it depends on the definitions and time-scale used. Cole and Cheshire (1996) presented a *mobile cellular automata* (MCA) model of ant colonies, looking at the interactions between active and inactive ants, and considered effects of the colony size, time of day, etc. using Fourier analysis to look for periodicity.

Many of these ideas are interesting, but it is not obvious how they could be applied to data such as the cow feeding data. The theories are more obviously applicable to situations such as the foraging behaviour and intake composition of large groups of animals, whereas I am more interested in modelling the individual animals and, to begin with at least, only a single food type. For choice cows it would be possible to investigate whether their relative intake of high- and low-protein food followed the IFD in some way; however with the relatively small number of animals and the complications of a social hierarchy, this is not an area that I will explore at the moment.

1.3.3 Stochastic systems, fractals and power laws

Behaviour data can be collected and recorded in different ways, for example as a point process, recording when events or behaviour changes occur in continuous time, or as recordings in discrete time, e.g. second by second, of which category of behaviour is being performed. Alados et al. (1996) considered examples of each of these data types, considering head-lifting behaviour in Spanish ibex and fitting simple linear models to frequencies on the log-log scale, the slope of which is defined to be the *fractal dimension.* This can be viewed as a measure of behavioural complexity, more complex behaviour being seen as advantageous to the animal. It is hypothesised that fractal dimension falls when an animal is disrupted or under stress, i.e. the slope of the log-log plots is smaller for the stressed animals. This reduction of dimension corresponds to a reduction in the animal's repertoire of behaviours, such as a stressed animal in a zoo exhibiting stereotypies. Evidence of a reduction in fractal dimension was found for animals that were under stress due to either pregnancy or parasitism.

An important feature of fractals and also *power laws* is the *scale invariance* property, the idea that the same laws hold on the macroscopic scale as on the microscopic scale. For example, time between earthquake tremors or sizes of avalanches may be described by power law distributions, the important thing to note being that long gaps between tremors or large avalanches are still part of the same distribution that describes the short gaps or small ones. The theory of *self-organised*

criticality (SOC), see for example Bak (1997), states that nature operates at a unique critical state. If we think about a scale of 'order', one end of the scale is that corresponding to perfect order, whilst the other is perfect randomness. Somewhere in between lies the *critical* state, where individual events are unpredictable, but the overall distribution of events is predictable, hence the critical state is considered the only 'interesting' state to be in. It should be remembered though that evidence of a power law type of relationship can be a simple consequence of events following a particular distribution and the fact that a power law relationship has been observed does not automatically imply SOC.

Ferriere et al. (1996) and Roberts (1994) both considered vigilance behaviour, looking at successive scan and inter-scan durations, similar to feeding data. Ferriere et al. (1996) presented ideas in terms of non-linear dynamical systems, looking for periodic patterns by considering Poincaré surfaces and Lyapunov exponents, whilst Roberts (1994) looked at inter-scan intervals and autocorrelations, finding that successive inter-scan intervals were largely unrelated, but that first differences of the series were predictable from the preceding few. An autoregressive model was fitted, resulting in coefficients being negative with decreasing magnitude. Checks were made via simulation and spectral analysis used to look for periodic components. It is not clear however that such ideas would be directly transferable to feeding behaviour. *Satiety* in feeding behaviour is the idea that once a meal has been finished, it will not be necessary to feed again for a while. This is not a feature of the vigilance behaviour discussed in these papers, being something that the animal is concerned with continuously and often involving cooperation between animals. Therefore again, although the methods discussed are mathematically interesting, this is not an area I have chosen to progress in relation to the cow feeding behaviour.

1.4 Some statistical models

I now describe some statistical models to consider as a starting point for the statistical modelling of categorical behaviour data. I discuss issues such as whether the models are based in continuous or discrete time and what types of latent structure are incorporated. Figure 1.3 shows how some of the models discussed below can be seen as generalisations of each other. An arrow indicates that the generalisation of a particular class of model leads to the other class. Some of the connections between them are discussed in the following sections. Hidden Markov and semi-Markov models are highlighted as models which are considered in depth

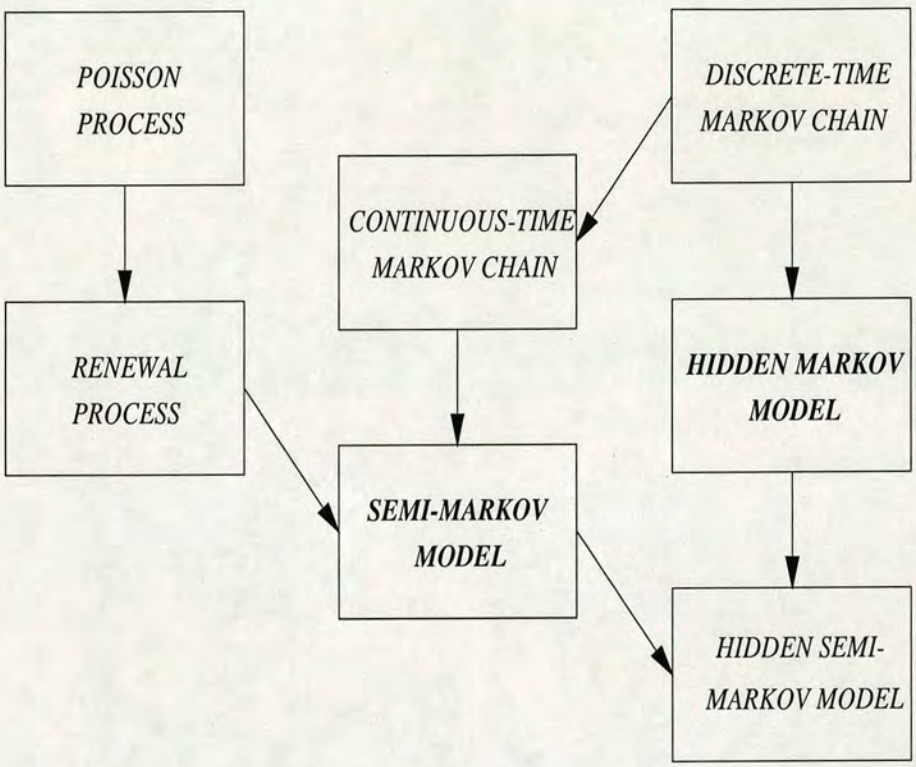

Figure 1.3: *Representation of relationships between types of model. Arrows indicate generalisation of models. Hidden Markov and semi-Markov models are highlighted as models which are considered in detail in later chapters.*

in subsequent chapters.

1.4.1 Continuous or discrete time

Data such as feeding data always occur naturally in continuous time. In its simplest form the data consist of just start- and end-times of feeding and so can be represented as in Table 1.1, where 0 indicates a non-feeding period and 1 a feeding period, each lasting for the given duration. The i-th event in such a series can be written (S_i, τ_i) for $i = 1, \ldots, N$, where S_i indicates the current category of behaviour and τ_i is the associated duration. In theory the durations are measurable to any precision; in practice they might be rounded to the nearest second or minute, say. Some models can only be formulated in discrete time and so it is useful to consider the alternative format given in Table 1.2, where we have arbitrarily chosen one minute as the timescale for discretisation, and the current behaviour being performed at each minute is recorded. For this representation we write the behaviour at time t as x_t for $t = 1, \ldots, n$. In such a framework, a rule must be chosen to categorise minutes in which the animal feeds for only part of that minute, the most obvious being to record a 1 if feeding occurred for more

Behaviour	*Duration (mins)*
0	108
1	9
0	19
1	40
0	441
1	114
0	144
1	.
.	.

Table 1.1: *Data recorded in continuous time;* 0 *is non-feeding,* 1 *is feeding.*

Time (mins)	*Behaviour*
1	0
2	0
3	0
.	.
.	.
108	0
109	1
110	1
.	.
.	.
117	1
118	0
119	0
.	.
.	.

Table 1.2: *Data recorded in discrete time;* 0 *is non-feeding,* 1 *is feeding.*

than half of that minute and a 0 otherwise.

It is important that a model in discrete time is invariant to the arbitrary timescale chosen, e.g. if the scale is doubled, the same model should be retained and parameters should correspond to the original model. Ideally it would also be desirable for the model to have an analogue in continuous time, as this is how the data originally occurred.

1.4.2 Simple compartment models

The simplest type of point process (Cox and Isham, 1980) is a *Poisson process*, consisting of events occurring randomly in time, inter-event times being exponential. A *compartment model* for feeding data might consist of alternating periods of time spent feeding and non-feeding, and durations in each state modelled by exponential distributions. This can also be considered a type of *marked Poisson process*, i.e. a Poisson process for which a real-valued random variable is attached to each point. Either the start of (non-)feeding periods can be considered as the events in the Poisson process with the durations attached as the marks, termed a *compound Poisson process*, or we can consider events occurring as one of two types, i.e. the start of a feeding event or the start of a non-feeding period, and consider a *multivariate Poisson process* or a marked Poisson process for which the mark is an indicator of the class of point.

Clearly this is a simple model and there may be features of the data that these models based on Poisson processes fail to capture. Firstly, inter-event times, or equivalently event durations, may not be well-described by simple exponential distributions. Also, not unrelated to this, if events do not occur independently in time then serial dependence needs to be built into the model.

1.4.3 Explicit state durations

A *renewal process* (Cox, 1970) is a generalisation of a Poisson process, in which intervals between events are described by some specified probability density function. An *alternating renewal process* is one which has two states, the interval between events depending on the type of event at the start of the interval, i.e. whether it is the start of a feeding event or the start of a non-feeding period. This type of process is a special case of a semi-Markov process (see below) for which there are only two possible states. Generalisation from a Poisson process to a

renewal process thus allows direct specification for the marginal distributions of the feeding and non-feeding durations.

1.4.4 Dependence

Markov models are considered in detail in Haccou and Meelis (1994). A Markov model is one for which we have a set of states and consider the probability of moving from one state to another. The Markov, or *lack of memory*, property is that the probability of transition to a given state depends only on the current state and not on past states. This is easily generalised to an *r-th order* Markov model if the probability of moving to a given state is dependent upon the last r states but none further back. A *semi-Markov* process is a generalisation of a Markov process for which inter-event times need no longer be exponential but follow some specified distribution; the renewal process is a one-state semi-Markov process and the alternating renewal process is the two-state version. The probability density function for the duration in a given state is dependent on the type of event at the beginning of the duration (i.e. start of non-feeding or start of feeding) and more generally can also depend on the type of event ending it, not applicable here since we are only dealing with two behaviours. Markov processes can be formulated in both discrete and continuous time. For a model in continuous time, if the duration of each event is ignored, the sequence of states visited follows a discrete Markov chain. This is also true for a semi-Markov model. These ideas are discussed in more detail in Chapter 6.

1.4.5 Discrete latent states

A *hidden Markov model* (HMM), as described in MacDonald and Zucchini (1997), is a model that has a series of underlying states following a Markov chain. These states are unobserved, the observations themselves being conditionally independent of each other and dependent only on the current underlying state. Note that HMMs can only be formulated in discrete time, but nevertheless are still biologically attractive because we can think about modelling the underlying state of the animal, rather than the observed behaviour. For example we will see that for the feeding data we will be able to have a state that is nominally 'feeding' but which also allows for the short non-feeding periods when a cow is moving between feeders. This is an example of a situation for which a change in behaviour is observed, but there is perhaps no change in the underlying state of the animal, the observed behaviour change being simply a consequence of the physical na-

ture of feeding. HMMs are mathematically attractive too, because unlike models with more general dependencies, the likelihood can be written down explicitly and is of a straightforward form, simply being a product of matrices of transition probabilities and marginal probability density functions.

The Markov nature of this model dictates that the distribution of run-lengths of any particular type of behaviour, i.e. the durations of that behaviour, are constrained to follow a mixture of geometric distributions (the discrete-time analogue of exponential distributions). Therefore a useful generalisation is to consider a *hidden semi-Markov model* (HSMM) in order to allow marginal distributions to be specified directly. Such a model would capture both the appropriate marginal distributions and the dependence structure of the data. The use of this type of model is demonstrated by Sansom (1999) for the modelling of rainfall data. He considered a three-tier model, taking the form of 'events' separated by inter-event periods; within events there were 'shower' or 'rain' episodes, and within episodes there were the observed wet and dry periods. This is an appealing model, and applied to the feeding data would not need to be as complex. However the amount of computation involved in estimating parameters in a HSMM is considerably more than in the HMM and this prevents them being more widely used. Hidden Markov models are the subject of Chapter 5.

1.4.6 A continuous latent variable

Considered as a binary time series, feeding data is far from Gaussian. However, because of the nice properties of Gaussian variables, it would be desirable to find a transformation under which the data achieves approximate normality, thus allowing some of the many results derived for Gaussian processes to be used. Previously, two-stage methods have sometimes been employed, i.e. take a binary process and apply some distribution to the feeding periods, but a more coherent approach, developed by Glasbey and Nevison (1997) for rainfall data, is to apply a monotonic transformation to the data to achieve marginal normality. This defines a latent Gaussian variable, with zero rainfall corresponding to censored values below a threshold. The idea here would be to create an unobservable normally-distributed variable from our data for which periods of feeding correspond to the variable exceeding some threshold. Biologically, it does not seem unreasonable to consider the latent variable as representative of some physiological or neurological states within the animal which affect its motivation to feed. The levels of each of these will vary, but perhaps only when they all gain some threshold is the animal motivated to begin feeding again. This is an idea which, as far as we know, has

not been applied to animal behaviour. The background is further discussed and models developed in Chapter 3.

1.5 Structure of the thesis

In this introductory chapter, I outlined the motivation for the thesis and discussed the cow feeding data that will be considered throughout the thesis to motivate and illustrate the models considered. I also presented a brief review of the animal behaviour literature and outlined some statistical models that will form the basis for the models developed later. In Chapter 2, I look at some exploratory data analysis and modelling techniques, which give more insight into what types of model are appropriate for more detailed study. I investigate the marginal distributions of behavioural events, the stationarity of the data overall, and the extent of dependence and trend in the data. Chapters 3, 5 and 6 discuss the three main approaches to the modelling of the feeding data that I consider in detail. Chapter 3 looks at a continuous latent variable model, a model that considers the binary data to have arisen from the thresholding of an underlying Gaussian variable; Chapter 4 contains proofs relating to the representation of a Gaussian process in its spectral form. Chapter 5 looks at models that have a categorical latent variable, most notably hidden Markov models. I also consider discrete-time compartment models as a special case of hidden Markov models. In Chapter 6, I discuss semi-Markov models and techniques for model estimation when the underlying states are unknown. Chapter 7 brings together the relative merits of the three main approaches and discusses their fundamental differences. I also consider techniques for the comparison of non-nested models and develop a parametric bootstrap approach that can be used to assess the relative fit of the models. Finally, Chapter 8 forms a short review and discusses conclusions and ideas for further work.

Chapter 2

Exploratory modelling

In this chapter I look at exploratory data analysis and some preliminary approaches to modelling. In Section 2.1, I justify the treatment of the data as binary time series rather than focusing on intake, the more biologically obvious variable, and consider the marginal distributions of these variables. Section 2.2 looks at the marginal distributions of non-feeding durations and at how these can be used to split the feeding events into bouts. In Section 2.3, CUSUMs are used to look for overall trend in the data; Section 2.4 investigates the extent of serial dependency and how it may be modelled. Finally, in Section 2.5, the variance and correlation structure of the data is used to look for a *critical timescale* that corresponds to the longest time over which feedback is operating.

The cow feeding data were introduced in Section 1.2, two days of binary time series for Cow 41 being shown in Figure 1.2. All the data for the cows on the high-protein feed are shown similarly in Appendix A, along with four cows on each of the low-protein and choice diets.

2.1 Intake and feeding duration

I first explore the relationship between intake and feeding time and then go on to consider the marginal distributions of feeding durations.

2.1.1 Correlation between intake and feeding duration

Figure 2.1 shows, for two of the cows, the high correlation between the duration of feeding events and the associated intake. Table 2.1 shows the correlation coefficients for all eight high-protein cows. All are highly significant with $p < 0.001$.

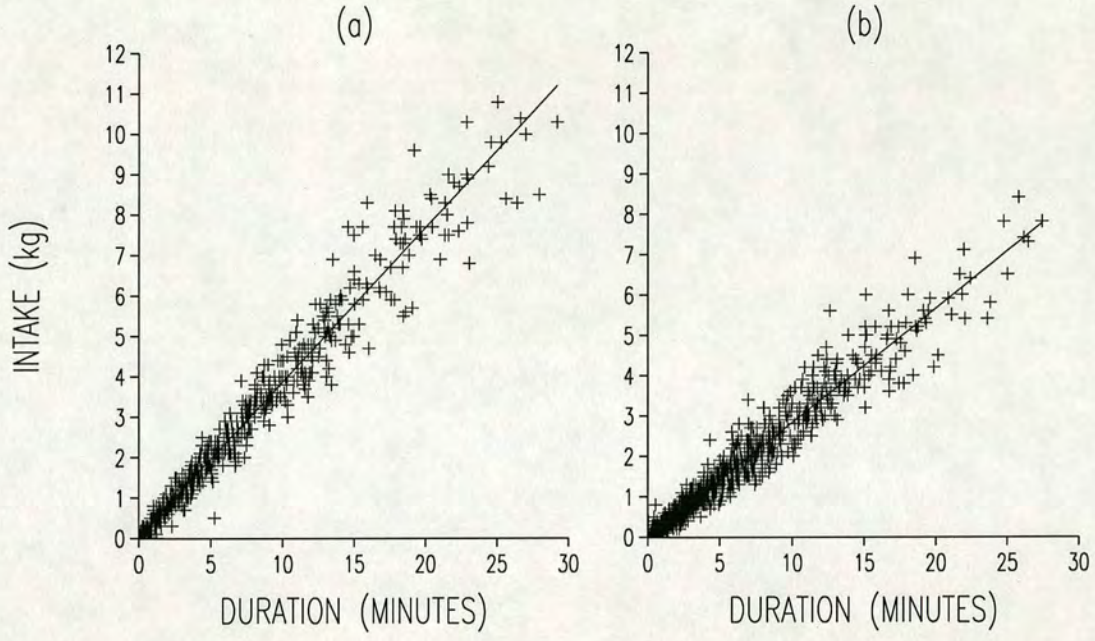

Figure 2.1: *Relationship between time spent at the feeder and amount of feed consumed for two of the high-protein cows; (a) Cow 5, (b) Cow 108.*

Therefore instead of modelling intake directly, which might seem the more biologically obvious thing to do, I concentrate on modelling time spent feeding, allowing the problem to be considered purely as one in time. It makes the assumptions that cows feed continuously whilst in the feeder and with constant rate. Given the high correlation between intake and feeding duration, these assumptions do not seem unreasonable, though there is no obvious way to further check them.

Cow	*Correlation*
5	0.981
41	0.956
108	0.974
169	0.950
170	0.949
182	0.957
194	0.967
221	0.971

Table 2.1: *Correlation between feeding duration and intake for the eight high-protein cows. All correlations are significant with p-values* < 0.001.

Cow	*N*	$\hat{\lambda}$	*Deviance*	*df*	*p-value*
5	587	0.1413	29.43	22	0.133
41	730	0.1838	42.95	25	0.014
108	944	0.1856	46.71	29	0.020
169	504	0.1557	38.01	20	0.009
170	897	0.2181	39.87	28	0.068
182	683	0.1627	46.10	24	0.004
194	771	0.1899	40.17	26	0.038
221	1323	0.2034	59.45	34	0.004
Pooled	6439	0.1823	203.38	78	< 0.001

Table 2.2: *Parameter estimates,* $\hat{\lambda}$*, and goodness of fit statistics for the fitting of exponential distributions to feeding durations for the eight high-protein cows. N is the total number of non-feeding events for each cow.*

2.1.2 Marginal distributions

Firstly, the histograms of feeding event duration and intake per feeding event are compared. We have already seen that they are highly correlated, and Figure 2.2 confirms that the two quantities have very similar marginal distributions. The figure shows, for each quantity, both a simple histogram and a log-transformed frequency plot, the approximate straight lines of the latter indicating that the marginal distributions are adequately described by exponential distributions. The plots are for the pooled data from all eight high-protein cows. Table 2.2 shows the fit of exponential distributions to feeding durations (in minutes) for individual cows and for pooled data. Maximum likelihood estimates of parameters, $\hat{\lambda}$, obtained in Genstat (Lawes Agricultural Trust, 1993), are given, along with the deviance, which can be used to test for evidence against the fitted distribution, having an asymptotic chi-squared distribution with the specified degrees of freedom. The statistics show evidence of lack of fit, but it should be remembered that for big datasets such as these, any small amount of lack of fit can show up in a significance test. Therefore even when the fit appears visually to be good, the significance test can dispute this. Figure 2.3 illustrates the fit for Cows 108 and 170, with parameters as given in Table 2.2. For both cows, and indeed for all eight cows, inspection of plots shows the fit to be good, even though the statistics of Table 2.2 disagree with this. I conclude that the exponential distribution does provide an adequate description of the marginal distributions of feeding durations, choosing to base conclusions on inspection of the plots rather than on the significance tests.

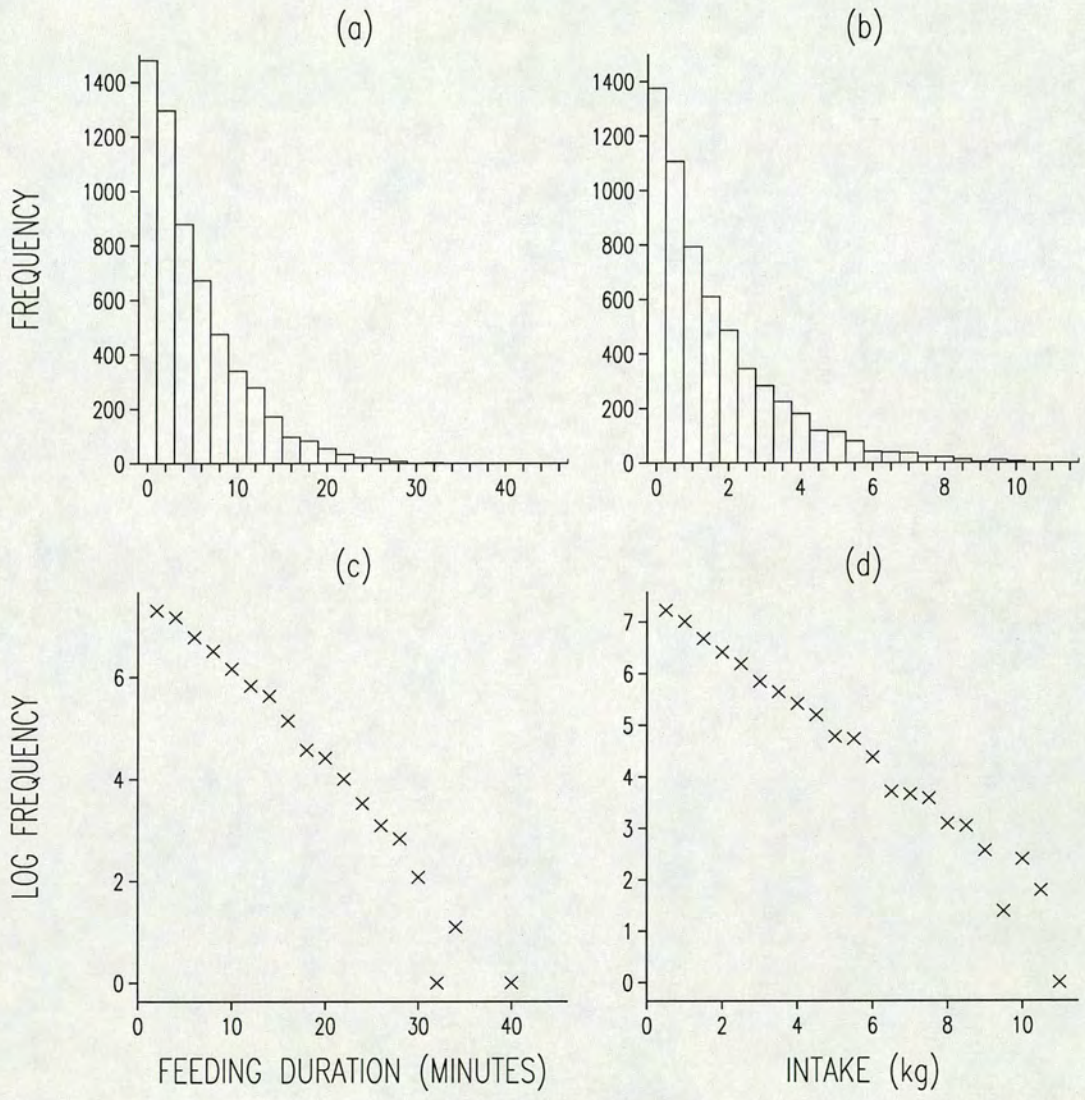

Figure 2.2: *Distributions of feeding durations and intake, based on pooled data from the eight high-protein cows; (a) histogram of feeding durations, (b) histogram of intake per feeding event, (c) log-frequency plot of feeding durations, (d) log-frequency plot of intake per feeding event.*

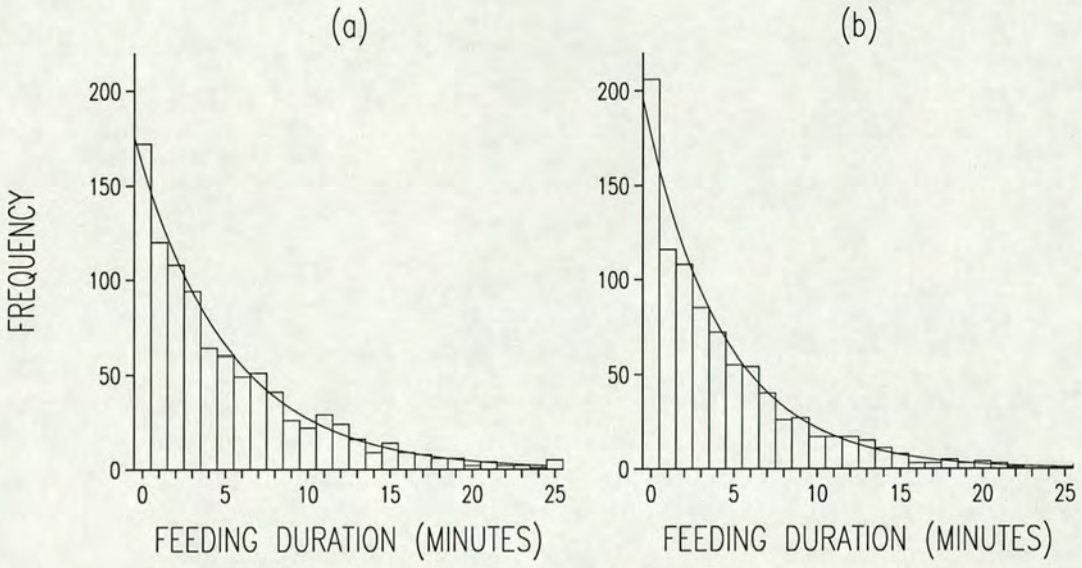

Figure 2.3: *Fit of exponential distributions to feeding durations; (a) Cow 108, (b) Cow 170.*

2.2 Non-feeding durations: splitting behaviour into bouts

If feeding events occurred randomly in time then inter-feed times would be exponentially distributed, the probability of feeding being constant no matter how long since the last feed, hence feeding would not be concentrated into bouts. It has already been noted that this is not the case. Feeding events are usually separated by short non-feeding periods and then groups of feeding events are separated by longer non-feeding periods. I first consider the marginal distributions of the non-feeding durations, and see how these can be used to split the feeding events into meals or bouts. In Section 1.2 it was discussed how it may be preferable to model feeding bouts instead of individual feeding events. If this is the desired approach, a criterion must be estimated to establish how long a non-feeding event must be in order for it to be considered as one that separates meals. This is called the *meal criterion* or *bout criterion.* Any non-feeding duration longer than this is then considered as between-meal and anything shorter as within-meal. We can then go on to group the individual feeding events into meals and if desired, think about frequency, duration, etc. of whole meals rather than of individual feeding events.

2.2.1 Meal criteria based on mixtures of exponential distributions

Many papers have looked at the issue of fitting distributions to the durations between occurrences of behaviour in order to divide the behaviour into bouts. Most, for example Fagen and Young (1978) and Slater and Lester (1982), plot log-survivorship functions and, assuming a mixture of two exponential distributions, fit a broken stick model and estimate the meal criterion as the breakpoint in the stick. Berdoy (1993) went on to consider a mixture of three exponentials instead of just two, and Langton et al. (1995) gave a much more thorough statistical treatment of the subject, using maximum likelihood to fit parameters, and carried out a simulation study to examine the behaviour of the likelihood ratio test statistic to decide between two- and three-process models.

To fit a mixture of two exponential distributions by maximum likelihood we maximise the log-likelihood given by

$$\sum_{i=1}^{N} \log\left(p\lambda_1 e^{-\lambda_1 \tau_i} + (1-p)\lambda_2 e^{-\lambda_2 \tau_i}\right),$$

where $\tau_i, i = 1, \ldots, N$, are the non-feeding durations, λ_1 and λ_2 are the parameters for the two exponential processes, sometimes labelled *fast* and *slow* respectively, and p is the proportion of events in the first (within-meal) distribution. As already mentioned, various criteria have been suggested as candidates for separating the two distributions. I briefly discuss three of them here.

Firstly, Fagen and Young (1978) and Slater and Lester (1982) minimise the total amount of time misclassified, giving a criterion $\hat{T}_1$, which corresponds to the crossing point of the two survivor functions. If plotted on the log-scale, these are straight lines and this can be thought of as a *broken stick* model. The criterion is given by the solution of

$$\hat{p}e^{-\hat{\lambda}_1 \hat{T}_1} = (1-\hat{p})e^{-\hat{\lambda}_2 \hat{T}_1},$$

$$\text{i.e. } \hat{T}_1 = \frac{1}{\hat{\lambda}_1 - \hat{\lambda}_2} \log\left(\frac{\hat{p}}{1-\hat{p}}\right).$$

A second option is to minimise the total number of events misclassified (Slater and Lester, 1982). This gives a criterion $\hat{T}_2$, which corresponds to the point at which the two probability density functions cross. This is calculated as the solution of

$$\hat{p}\hat{\lambda}_1 e^{-\hat{\lambda}_1 \hat{T}_2} = (1-\hat{p})\hat{\lambda}_2 e^{-\hat{\lambda}_2 \hat{T}_2},$$

$$\text{i.e. } \hat{T}_2 = \frac{1}{\hat{\lambda}_1 - \hat{\lambda}_2} \log\left(\frac{\hat{p}\hat{\lambda}_1}{(1-\hat{p})\hat{\lambda}_2}\right).$$

Cow	$\hat{p}$	$\hat{\lambda}_1^{-1}$	$\hat{\lambda}_2^{-1}$	$\hat{T}_1$	$\hat{T}_2$	$\hat{T}_3$
5	0.6278	1.607	174.1	0.85	8.45	6.22
41	0.7058	2.172	176.6	1.93	11.60	8.54
108	0.6588	0.789	116.4	0.52	4.49	3.33
169	0.5926	1.454	189.7	0.55	7.69	5.67
170	0.7498	1.515	167.8	1.68	8.88	6.60
182	0.7197	2.020	195.7	1.92	11.26	8.32
194	0.6432	0.999	139.6	0.59	5.57	4.12
221	0.7697	1.065	116.1	1.30	6.34	4.72
Pooled	0.6984	1.364	155.2	1.16	7.67	5.68

Table 2.3: *Estimates of parameters and meal criteria (in minutes) for mixtures of two exponential distributions fit to non-feeding durations for the eight high-protein cows.*

Finally, a third criterion is that which aims to misclassify equal numbers of events from each of the two distributions. This is calculated by equating the expected number of misclassifications from each distribution. The resulting criterion, $\hat{T}_3$, is given by the solution of

$$\hat{p}_1 e^{-\hat{\lambda}_1 \hat{T}_3} = \hat{p}_2 (1 - e^{-\hat{\lambda}_2 \hat{T}_3}).$$

This has no closed form solution but can be solved by an iterative process such as the Newton-Raphson method.

Whichever meal criterion $\hat{T}$ is used, the number of events considered to be misassigned, i.e. allocated to the wrong distribution, is

$$N\left[\int_{\hat{T}}^{\infty} \hat{p}\hat{\lambda}_1 e^{-\hat{\lambda}_1 \tau} d\tau + \int_0^{\hat{T}} (1-\hat{p})\hat{\lambda}_2 e^{-\hat{\lambda}_2 \tau} d\tau\right] = N\left[\hat{p}e^{-\hat{\lambda}_1 \hat{T}} + (1-\hat{p})(1 - e^{-\hat{\lambda}_2 \hat{T}})\right],$$

where N is the total number of non-feeding events.

Table 2.3 shows parameter estimates for mixtures of two exponential distributions fit to data from the eight high-protein cows. Likelihoods were maximised numerically using the FITNONLINEAR directive in Genstat (Lawes Agricultural Trust, 1993). Also shown are the resulting meal criteria, $\hat{T}_1$, $\hat{T}_2$ and $\hat{T}_3$, as defined above. Figure 2.4 displays the fit of the fitted distributions for Cows 5 and 108. It can be seen that the fit is poor. The histograms are bimodal, yet the density function for a mixture of exponentials is always decreasing (as the derivative of a mixture of exponential functions is always negative). Therefore even with more components, the bimodality of Figure 2.4 can never be captured. Further, we can estimate from the histograms that, for these two cows at least, the meal criterion should be somewhere around 30–60 minutes, corresponding to the observed troughs. So from these plots, and from what we would expect biologically, all

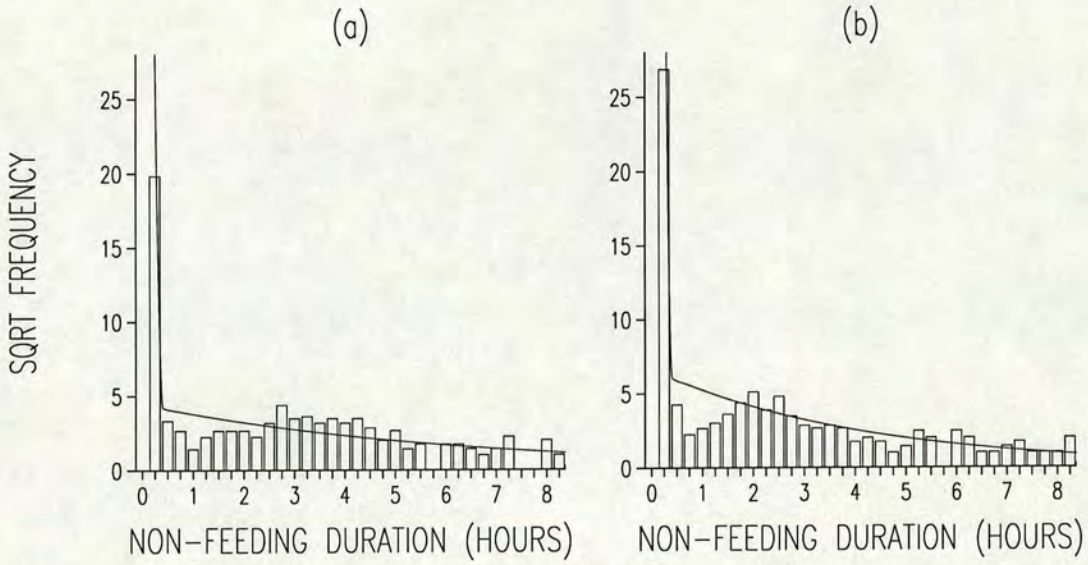

Figure 2.4: *Fit of mixtures of two exponential distributions to non-feeding durations; (a) Cow 5, (b) Cow 108. Frequencies are plotted on a square-root scale.*

three types of estimated meal criteria in Table 2.3 can be seen to be much too small.

A similar picture is obtained using data from the sixteen choice cows. For the pooled data, maximum likelihood parameter estimates are $\hat{p} = 0.7433, \hat{\lambda}_1 = 1.426, \hat{\lambda}_2 = 170.8$. From these estimates the three meal criteria can be calculated to be 1.53 minutes, 8.41 minutes and 6.26 minutes, respectively. These are very similar results to the high-protein cows discussed above, and inspection of similar plots (not shown) indicate that for these animals also, a realistic meal criterion should be at least 20 minutes.

2.2.2 Mixtures of log-normal distributions

It is clear that mixtures of exponential distributions do not describe the distribution of non-feeding durations well and hence we seek alternative distributions. We have already seen that the distributions are highly skewed, and so it is useful to consider the histogram of log-transformed durations. Figure 2.5 shows such histograms of pooled data for both the cows on the high-protein and choice diets. We already know that a single unimodal distribution cannot describe these histograms, and the shape here suggests that for durations on the log-scale, a mixture of normal distributions might be appropriate. This is equivalent to fitting mixtures of log-normal distributions to the untransformed data. Tolkamp

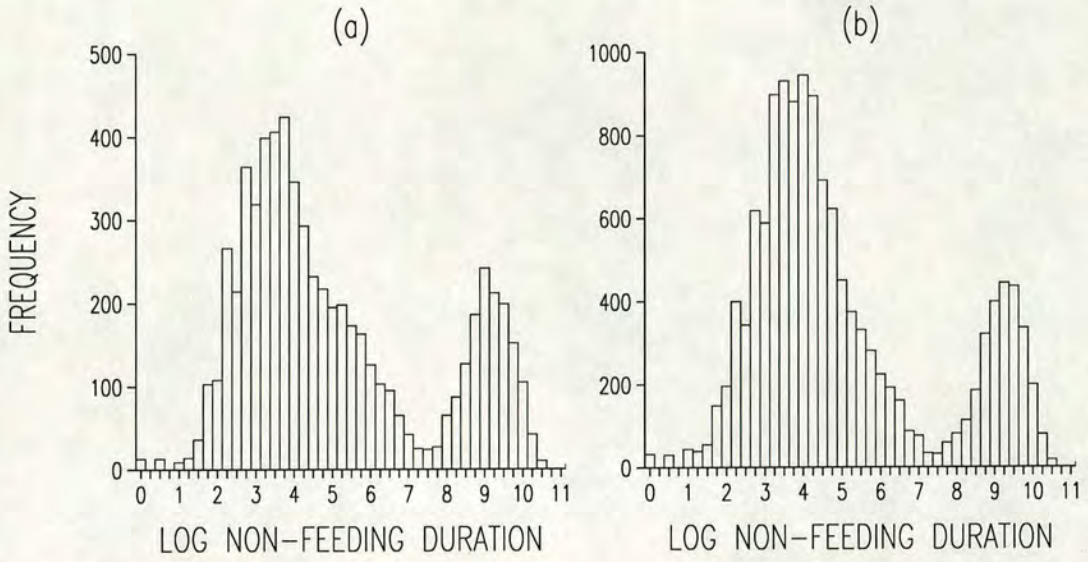

Figure 2.5: *Histograms of log-transformed non-feeding durations, based on pooled data; (a) for the eight high-protein cows (6439 events), (b) for the sixteen choice cows (13294 events).*

et al. (1998a) describe such an approach; more background and related earlier literature is discussed below.

Figure 2.6 shows the relative fit of the mixtures of two exponential and log-normal distributions to the pooled data from the choice cows. This shows the mixture of log-normal distributions to be a much better fit than the exponential mixture, and illustrates the way in which the exponential mixture was estimating the meal criterion to be too small; the trough predicted by the fitted exponential mixture is at much too low a value, whereas the log-normal mixture has its trough in a much more plausible position. Figure 2.7 displays the same fit in terms of log-survivorship and cumulative frequency plots, again illustrating the superior fit of the log-normal model.

Figures 2.8, 2.9 and 2.10, show histograms for three individual cows, and it becomes apparent that for some cows, such as Cows 5 and 170, a mixture of two component distributions looks reasonable, whereas for others, such as Cow 108, a mixture of three components looks necessary. Fitting a mixture of two distributions can be interpreted as directly classifying the non-feeding periods as within- or between-meal. The occurrence of the third distribution for some cows has been investigated by Tolkamp and Kyriazakis (1999b) and found to be because sometimes a cow will go from feeder to feeder via the drinking troughs at the end of the yard, resulting in a duration between feeding events that is longer than

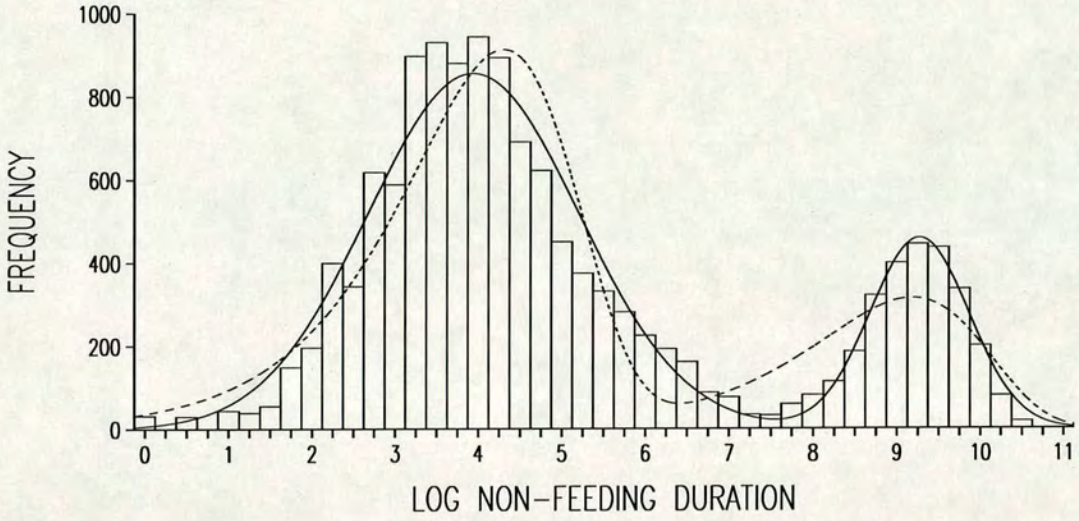

Figure 2.6: *Fit of mixture distributions for pooled data from the 16 choice cows; (—) mixture of two log-normal distributions, (- - -) mixture of two exponential distributions.*

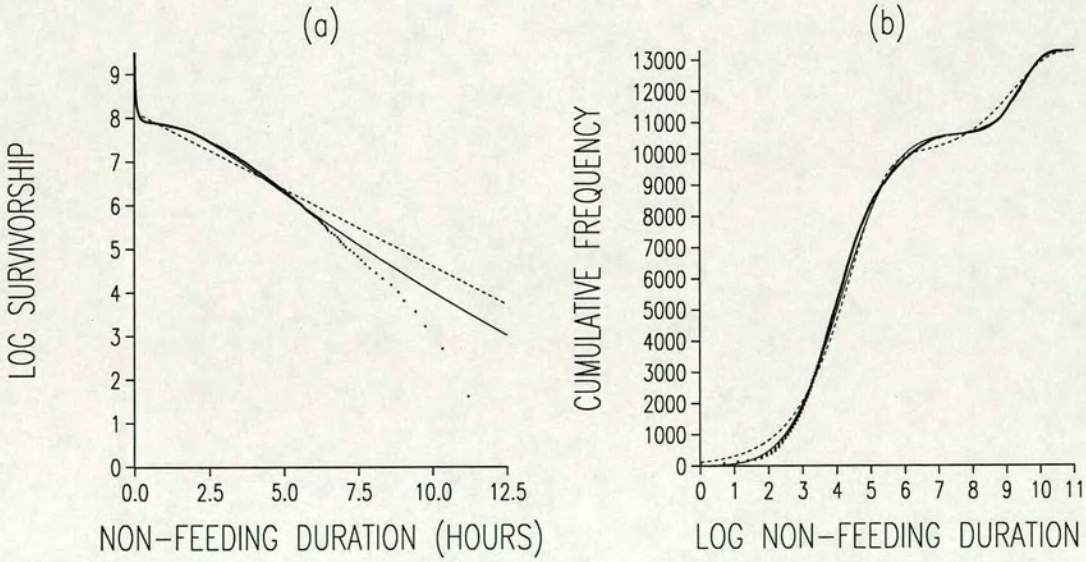

Figure 2.7: *Comparison of (—) mixture of two log-normal distributions, and (- - -) mixture of two exponential distributions, for pooled data from the sixteen choice cows (· · ·); (a) log-survivorship curve, (b) cumulative frequency plot.*

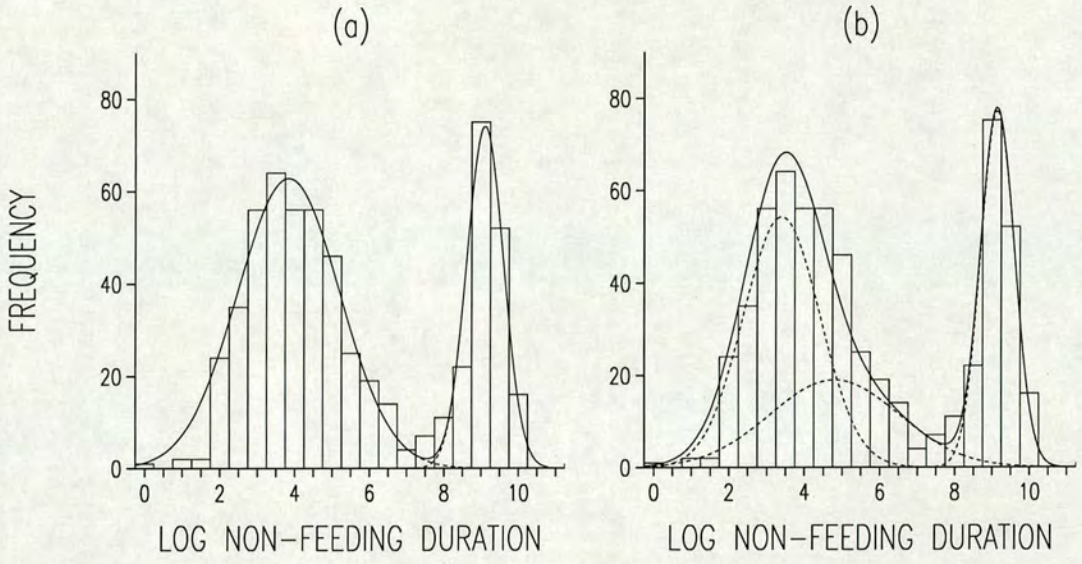

Figure 2.8: *Fit of log-normal mixture distributions to non-feeding durations for Cow 5; (a) mixture of two, (b) mixture of three distributions; (- - -) component distributions, (—) combined mixture distribution.*

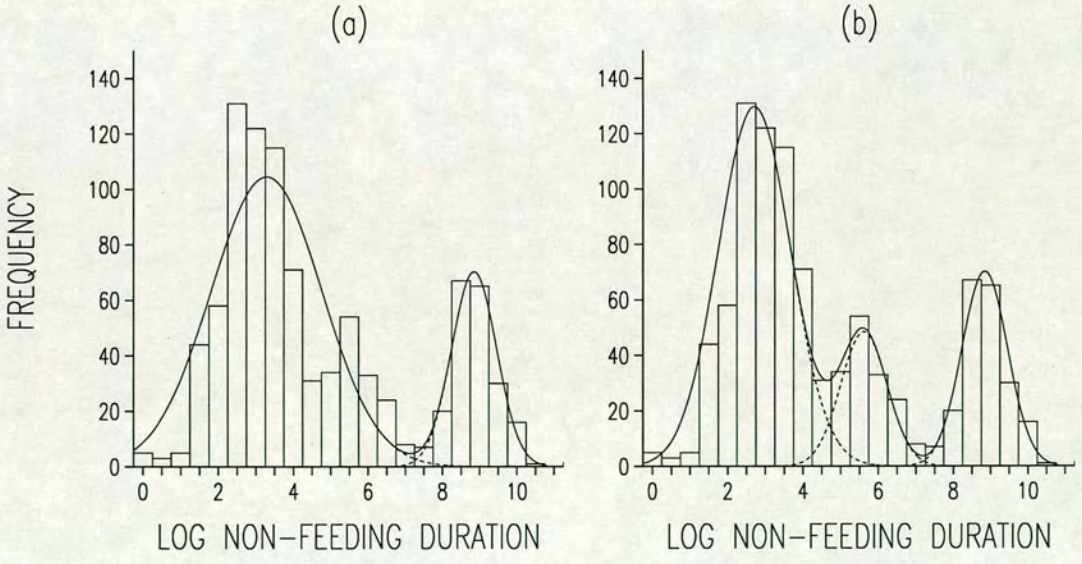

Figure 2.9: *Fit of log-normal mixture distributions to non-feeding durations for Cow 108; (a) mixture of two, (b) mixture of three distributions; (- - -) component distributions, (—) combined mixture distribution.*

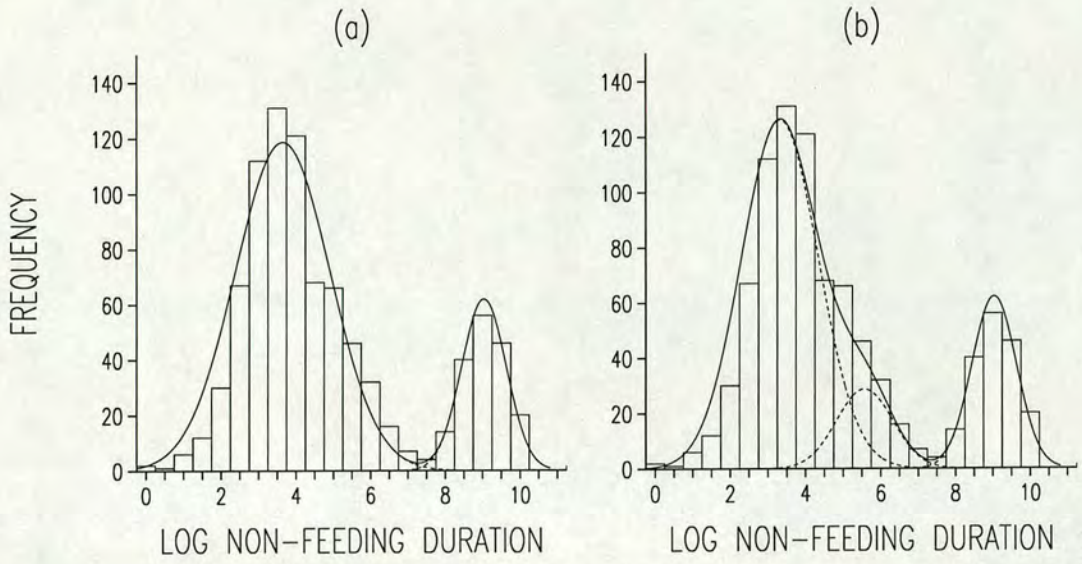

Figure 2.10: *Fit of log-normal mixture distributions to non-feeding durations for Cow 170; (a) mixture of two, (b) mixture of three distributions; (- - -) component distributions, (—) combined mixture distribution.*

a typical within-meal gap, but shorter than a between-meal gap. Therefore in going from a mixture of two to three distributions, the between-meal distribution remains very similar, and the within-meal distribution splits into two.

To fit a mixture of K log-normal distributions we maximise the log-likelihood given by

$$\mathcal{L}_K = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} p_k \frac{1}{\sqrt{2\pi}\sigma_k \tau_i} e^{-\frac{(\log \tau_i - \mu_k)^2}{2{\sigma_k}^2}} \right),$$

where $\tau_i, i = 1, \ldots, N$, are the non-feeding durations, $(\mu_k, {\sigma_k}^2)$ are the parameters for each component distribution, and p_k are the proportions of non-feeding durations estimated to come from each component distribution, with the constraint

$$\sum_{k=1}^{K} p_k = 1.$$

Results are presented for mixtures of both two and three log-normal distributions to non-feeding durations. Tables 2.4 and 2.5 show parameter estimates for the individual cows on the high-protein feed and for their pooled data. It is reassuring to note that for each cow the sum of the estimated proportions $\hat{p}_1$ and $\hat{p}_2$ in Table 2.5 is roughly equal to $\hat{p}$ in Table 2.4, confirming that the addition of the extra distribution is for description of the longer within-meal events, hence the estimates for the distribution of between-meal events are very similar in both models, i.e. compare $(\hat{\mu}_2, \hat{\sigma}_2)$ in Table 2.4 with $(\hat{\mu}_3, \hat{\sigma}_3)$ in Table 2.5.

Cow	$\hat{p}$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$
5	0.7014	4.310	1.307	9.351	0.472
41	0.7873	4.472	1.475	9.465	0.423
108	0.7841	3.783	1.412	9.108	0.579
169	0.6577	4.210	1.229	9.323	0.630
170	0.8019	4.125	1.209	9.281	0.572
182	0.8123	4.405	1.469	9.598	0.545
194	0.7158	3.792	1.321	9.038	0.660
221	0.8412	3.863	1.235	8.995	0.634
Pooled	0.7744	4.062	1.341	9.234	0.607

Table 2.4: *Parameter estimates for mixtures of two log-normal distributions fit to non-feeding durations for the eight high-protein cows.* $\hat{p}$ *is the estimated proportion of events in the distribution with estimated parameters* $(\hat{\mu}_1, \hat{\sigma}_1^2)$.

Cow	$\hat{p}_1$	$\hat{p}_2$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\mu}_3$	$\hat{\sigma}_3$
5	0.4514	0.2630	3.892	0.973	5.236	1.637	9.387	0.434
41	0.1620	0.6271	3.098	0.563	4.840	1.441	9.472	0.416
108	0.6301	0.1520	3.233	0.914	6.009	0.591	9.098	0.585
169	0.3301	0.3343	3.418	0.677	5.065	1.197	9.352	0.597
170	0.6869	0.1150	3.816	0.975	5.966	0.717	9.281	0.570
182	0.2776	0.5433	3.216	0.630	5.078	1.455	9.649	0.493
194	0.5475	0.1697	3.270	0.937	5.504	0.899	9.047	0.649
221	0.7590	0.0820	3.612	1.004	6.174	0.542	8.995	0.630
Pooled	0.5927	0.1822	3.528	0.953	5.806	0.857	9.239	0.600

Table 2.5: *Parameter estimates for mixtures of three log-normal distributions fit to non-feeding durations for the eight high-protein cows.* $\hat{p}_1$ *is the estimated proportion of events in the distribution with estimated parameters* $(\hat{\mu}_1, \hat{\sigma}_1^2)$*;* $\hat{p}_2$ *is the corresponding proportion for the distribution with* $(\hat{\mu}_2, \hat{\sigma}_2^2)$. $\hat{p}_3$ *is given by* $(1 - \hat{p}_1 - \hat{p}_2)$.

Table 2.6 shows statistics to enable the comparison of the models to see whether there is convincing evidence to include the third distribution. It might be first thought that these models can be simply compared using the likelihood ratio statistic, however testing between different numbers of components in a mixture model is equivalent to testing whether one of the mixing parameters is equal to zero. As this is the smallest value it can take, i.e. on the boundary of the parameter space, the asymptotic results needed for the theory of the likelihood ratio test are invalid, see for example McLachlan (1987) or Titterington (1990). An alternative approach is to consider criteria such as Akaike's or Schwarz's Bayesian information criteria (AIC and BIC, respectively), however these are also strictly affected by the lack of validity of the asymptotic results, but we present these in Table 2.6 due to the lack of any simple alternative. A better approach would be to use a parametric bootstrap, simulating from the two-component mixture and then bootstrapping the likelihood ratio statistic. However the choice between a two- and three-component mixture model, whilst forming an interesting biological problem, is not crucial for the models we will be developing and therefore we quote simply AIC and BIC, given by

$$\begin{aligned} \mathrm{AIC}_K &= -2\mathcal{L}_K + 2m_K \\ \mathrm{BIC}_K &= -2\mathcal{L}_K + m_K \log N \end{aligned}$$

where $\mathcal{L}_K$ is the maximised log-likelihood of the K-component mixture, N is the number of events and m_K is the number of parameters estimated. Here we have $m_2 = 5$ and $m_3 = 8$. Both are penalised likelihood tests, i.e. they can be thought of as the log-likelihood modified by a penalty for the number of parameters estimated. The practice is to select the model which gives the lowest value for the preferred criterion.

It can be seen that AIC favours the three-component model in all cases; BIC gives almost the same conclusions, except Cow 5, for which the two-component model gives a slightly lower value. For some of the other cows BIC gives similar values for both models. These criteria should not be considered in isolation, it being important to always inspect the fit of the models visually. For Cow 5, the result here confirms the mixture of two distributions to be an adequate fit, as already noted from Figure 2.8. For Cow 108, Figure 2.9 showed clear evidence of a third distribution; this is confirmed by the relatively large increase in likelihood when the third distribution is included and the lower values of AIC and BIC for the three-component mixture. For Cow 170 there is little reduction in likelihood and a correspondingly small decrease in AIC; BIC gives the same value for both models. Hence it seems there is little to be gained by using a three-component

Cow	N	$-\mathcal{L}_2$	$-\mathcal{L}_3$	AIC_2	AIC_3	BIC_2	BIC_3
5	587	4575	4567	9161	9150	9183	9185
41	730	5531	5516	11072	11049	11095	11086
108	944	6621	6560	13253	13136	13277	13175
169	504	4024	4007	8057	8030	8078	8064
170	897	6365	6355	12741	12726	12765	12764
182	683	5094	5067	10198	10149	10221	10185
194	771	5680	5664	11369	11344	11392	11381
221	1323	8772	8741	17554	17497	17580	17539
Pooled	6439	46888	46752	93786	93520	93820	93574

Table 2.6: *Comparison of mixtures of two and three log-normal distributions fit to non-feeding durations for the eight high-protein cows. N is the total number of observations for each cow, $\mathcal{L}_K$ is the log-likelihood for the K-component mixture and AIC_K and BIC_K are the corresponding information criteria.*

mixture for this cow, as already noted from Figure 2.10. Therefore it is clear that decisions about which is the best model should be made on an individual basis and should include inspection of the relevant plots.

2.2.3 Meal criteria based on mixtures of log-normal distributions

As previously, different definitions of meal criteria are available, the obvious two here being either to minimise the total number of events misclassified or to misclassify the same expected number from both distributions. These are given by $\hat{T}_a$ and $\hat{T}_b$ respectively, as the solutions to the equations

$$\hat{p}_k \, \phi\left(\frac{\log \hat{T}_a - \hat{\mu}_k}{\hat{\sigma}_k}\right) = \hat{p}_l \, \phi\left(\frac{\log \hat{T}_a - \hat{\mu}_l}{\hat{\sigma}_l}\right)$$

and

$$\hat{p}_k \left(1 - \Phi\left(\frac{\log \hat{T}_b - \hat{\mu}_k}{\hat{\sigma}_k}\right)\right) = \hat{p}_l \, \Phi\left(\frac{\log \hat{T}_b - \hat{\mu}_l}{\hat{\sigma}_l}\right)$$

where $\phi(z)$ is the density function of the standard normal distribution, and $\Phi(z)$ the distribution function, i.e. $P(Z \leq z)$. For a two-component model, $k = 1$ and $l = 2$; for a three-component mixture we can use $k = 2$ and $l = 3$, as the distributions are sufficiently well separated to ignore the contribution from the first distribution in this case. The good separation also means the two definitions of meal criteria produce very similar estimates. Below we use only the criterion $\hat{T}_a$, which can be calculated as the solution to a quadratic equation; criterion $\hat{T}_b$ has to be calculated numerically.

Cow	N	*2 components*		*3 components*	
		$\log \hat{T}$	$\hat{T}$	$\log \hat{T}$	$\hat{T}$
5	587	7.994	**49.4**	8.296	66.8
41	730	8.360	**71.2**	8.415	75.3
108	944	7.621	34.0	7.520	**30.7**
169	504	7.588	**33.0**	7.811	41.2
170	897	7.714	**37.3**	7.718	37.5
182	683	8.267	64.9	8.496	**81.6**
194	771	7.329	**25.4**	7.425	28.0
221	1323	7.410	**27.6**	7.418	27.8
Pooled	6439	7.693	36.5	7.740	**38.3**

Table 2.7: *Estimated meal criteria for the high-protein cows based on mixtures of two and three log-normal distributions fit to non-feeding durations. N is the total number of events for each cow, $\log \hat{T}$ is the criterion as calculated in log-seconds and $\hat{T}$ is the back-transformed value in minutes. The criteria judged to be the more appropriate is highlighted.*

Whichever criterion, $\hat{T}$, is chosen, the expected number of events assigned to the wrong distribution can be calculated as

$$\begin{aligned} & N\left[\int_{\log \hat{T}}^{\infty} \hat{p}_k \, \phi\left(\frac{\log \tau - \hat{\mu}_k}{\hat{\sigma}_k}\right) d\log\tau + \int_{-\infty}^{\log \hat{T}} \hat{p}_l \, \phi\left(\frac{\log \tau - \hat{\mu}_l}{\hat{\sigma}_l}\right) d\log\tau\right] \\ = \; & N\left[\hat{p}_k\left(1 - \Phi\left(\frac{\log \hat{T} - \hat{\mu}_k}{\hat{\sigma}_k}\right)\right) + \hat{p}_l \, \Phi\left(\frac{\log \hat{T} - \hat{\mu}_l}{\hat{\sigma}_l}\right)\right], \end{aligned}$$

where N is the total number of non-feeding events, and $\phi(\cdot)$ and $\Phi(\cdot)$ are as defined above.

I now compare the meal criteria from fitting the two- and three-component mixtures. Table 2.7 gives individual and pooled meal criteria for these two models for the eight high-protein cows. In conjunction with these criteria, again it is essential to inspect the histograms for individual cows, as in Figures 2.8–2.10. We have already noted that for Cow 108 there is strong evidence of a third distribution, for Cow 170 any evidence is much weaker and for Cow 5 there is little evidence at all.

Although the number of distributions appropriate for the mixture is of interest from the point of view of understanding the behavioural organisation of the cows, for the allocation of visits to meals and hence for the splitting up of feeding events into bouts, there is often little difference in results from use of either set of criteria as shown in Table 2.7. This is due to the good separation between the distributions and hence the small number of non-feeding events with duration around the meal criteria. Nevertheless I show in the table which of the criteria is preferable for each cow. This is based on whether there is evidence for a third

distribution, using both histograms and the statistics in Table 2.6, and on which value of the meal criteria is more appropriate.

Briefly returning to the sixteen choice cows, from the pooled data, maximum likelihood estimates for the parameters for the two-component mixture are $\hat{p} = 0.800, \hat{\mu}_1 = 4.099, \hat{\sigma}_1 = 1.236, \hat{\mu}_2 = 9.313$ and $\hat{\sigma}_2 = 0.578$. The means for the two distributions on the absolute time scale are 129 seconds and 218 minutes respectively, and the meal criterion is calculated as 38.3 minutes, corresponding to a value of 7.7 on the log-seconds scale of Figure 2.6. This value looks perfectly reasonable as the separation point between the two distributions; recall that the meal criteria resulting from the mixture of exponential distributions were much too low.

2.2.4 Discussion on meal criteria

It was discussed above how papers have previously modelled non-feeding durations with a mixture of exponential distributions. From the last section, and as found in Tolkamp et al. (1998a) and Tolkamp and Kyriazakis (1999a), we have seen that a better alternative is to use a mixture of log-normal distributions. From Figure 2.4 we saw that a mixture of exponential distributions could not adequately describe the observed shape of the data and we found that meal criteria were estimated too short. For the estimation of meal criteria it is particularly important to obtain a good fit in the region between the two components of the mixture, and from Figure 2.6 particularly, the log-normal mixture is seen to far out-perform the exponential model in this respect.

The bimodality of the histograms can be linked to the concept of satiety, whereby after a meal, cows go through a period of not wanting to eat, as their hunger is still satisfied from their previous meal. Therefore the lack of memory associated with exponential inter-feeding times is not a sound assumption to make. Considering a cow's motivation to begin feeding as a consequence of the states of various physiological/neurological factors, each independent and each being required to reach a certain threshold to motivate feeding, the probability of feeding is the product of the probabilities of several individual conditions being satisfied, the log-normal distribution is a logical distribution to consider. Montroll and Shlesinger (1982) and Tolkamp et al. (1998a) give more discussion on this. Another alternative, suggested by Simpson and Ludlow (1986) and Yeates et al. (2001), is to use Weibull distributions to model inter-feed durations; they argue that its monotonically increasing hazard function is more biologically sound than

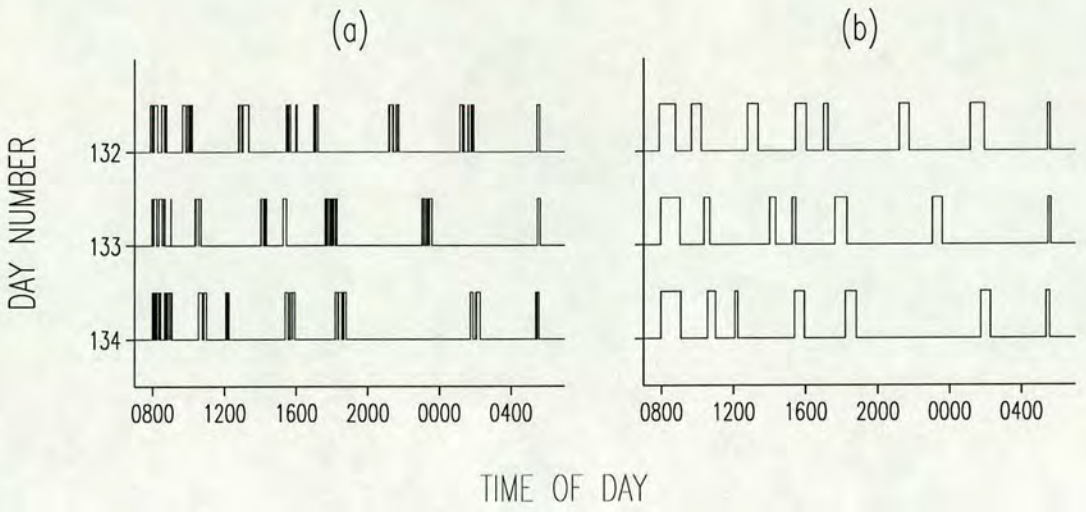

Figure 2.11: *Three days of data for Cow 108; (a) feeder-visit data; (b) meal data.*

that of the log-normal, which increases to a maximum before decreasing again. However within this thesis I accept the log-normal description as being adequate for our purposes. For later work, the particular choice of distribution does not affect the methodology developed, though I recognise that the Weibull distribution may be a better alternative.

Finally in this section we briefly discuss the analysis of meal data as opposed to feeder-visit data. Figure 2.11 shows example data for Cow 108, both individual feeding event data and after allocation of visits to meals, using a meal criterion of 30.7 minutes as given in Table 2.7. Figure 2.12 shows the marginal distribution of meal durations that results from the application of this meal criterion. It can be seen that on an absolute timescale, the distribution of meal durations is skewed to the right. After log-transformation, the distribution is more symmetric, but becomes slightly left-skewed, indicating that the data have been slightly over-transformed. This is typical for all the eight high-protein cows. Tolkamp et al. (2000) point out that for some objectives, meal data may be more biologically relevant to model than individual visit data, arguing that the latter can be influenced by the management regime under which the animals are kept, and therefore if the objective of modelling involves the comparison of animals kept under different regimes, meal data are more consistent to work with. However, as argued in Chapter 1, as the main objective here is to look at short-term behaviour, this includes intra-meal behaviour, and so in general we want to develop models that can explain Figure 2.11(a) rather than Figure 2.11(b). In most cases, a model that can adequately explain individual visit data has a simpler version that can explain the corresponding meal data.

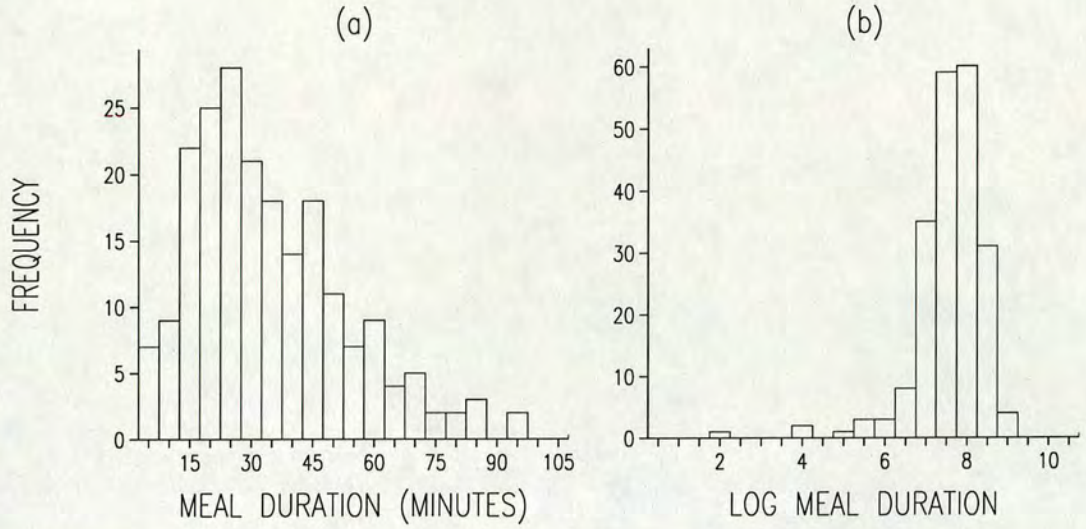

Figure 2.12: *Marginal distribution of meal sizes for Cow 108 based on a meal criterion of 30.7 minutes; (a) absolute timescale, (b) log-transformed timescale.*

2.3 Stationarity

The first assumption usually made in the analysis of time series data is that of stationarity, so I will first investigate whether there are any obvious overall trends in the data over the 30 days, in terms of both feeding and non-feeding durations. As well as looking at simple plots, I also consider CUSUM plots and their associated Kolmogorov-Smirnov test statistics. These use the overall rank of (non-)feeding durations to test for evidence that the order in which they occur is not random. Glasbey and Martin (1986) used such tests to detect changes in the behaviour of single ion channels, which have open and closed durations. For a series of durations $\tau_i, i = 1, \ldots, N$, the CUSUM plot is defined as the plot of cumulative sums S_l, given by

$$S_l = \sum_{i=1}^{l} (\tau_i - \bar{\tau}),$$

plotted against l for $l = 1, \ldots, N$, where $\bar{\tau}$ is the mean event-duration of the whole series.

If events occur in random order, the CUSUM would be expected to be close to zero. If calculations are based on ranks instead of actual durations, the standard Kolmogorov-Smirnov test statistic can be used to get a significance level for the deviation from zero. It is defined as

$$D = \max |S_l| \quad \text{for } l = 1, \ldots, N.$$

Cow	N	*Feeding*		*Non-feeding*	
		D	p	D	p
5	587	1800	1.000	2791	0.795
41	730	7952	0.041	3609	0.896
108	944	15151	0.003	5188	0.929
169	504	1076	1.000	2817	0.453
170	897	13778	0.004	19036	< 0.001
182	683	11051	< 0.001	8538	0.008
194	771	10434	0.007	8737	0.037
221	1323	14795	0.207	23105	0.008

Table 2.8: *Kolmogorov-Smirnov test statistics, D, and associated p-values for testing whether durations occur in random order.*

An approximate probability level for this statistic is then given by

$$\min\left(2\exp\left\{\frac{-24D^2}{N^2(N+1)}\right\},\ 1\right).$$

Plots for both feeding and non-feeding events were inspected for all cows on the high-protein feed. Figures 2.13 and 2.14 show plots for two of the cows, for both ranked feeding durations and ranked non-feeding durations. Cow 5 has $N = 587$ observations in each of the series; Cow 108 has $N = 944$. Table 2.8 shows the Kolmogorov-Smirnov statistics and their associated p-values for these plots, and includes results for all eight high-protein cows. For Cow 5 there is no evidence that the ranks are not in random order for either feeding or non-feeding events. For Cow 108 however, there is strong evidence that the feeding events are not in random order, $p = 0.003$, illustrated in Figure 2.14(b) by the CUSUM having a large departure from zero in the middle of the series. The shape of this plot indicates an excess of longer feeding durations near the beginning of the series compared with towards the end. From Table 2.8 we can see that for five of the eight cows there is evidence that the feeding durations are not in random order, and for four cows there is evidence that the non-feeding durations are not in random order. It is interesting to note that in all the cases for which there is evidence of the order being non-random, the plots are the same shape, as in Figure 2.14(b), so indicating an excess of longer feeding durations near the beginning. I have not come up with a viable biological explanation as to why this should be the case.

It should be noted that these results should be treated with some caution — if the ranks are in a random order this implies that the process is stationary; however the converse is not true, as if a stationary process has correlation between successive observations, as is suspected with this data, the ranks will not be in random order

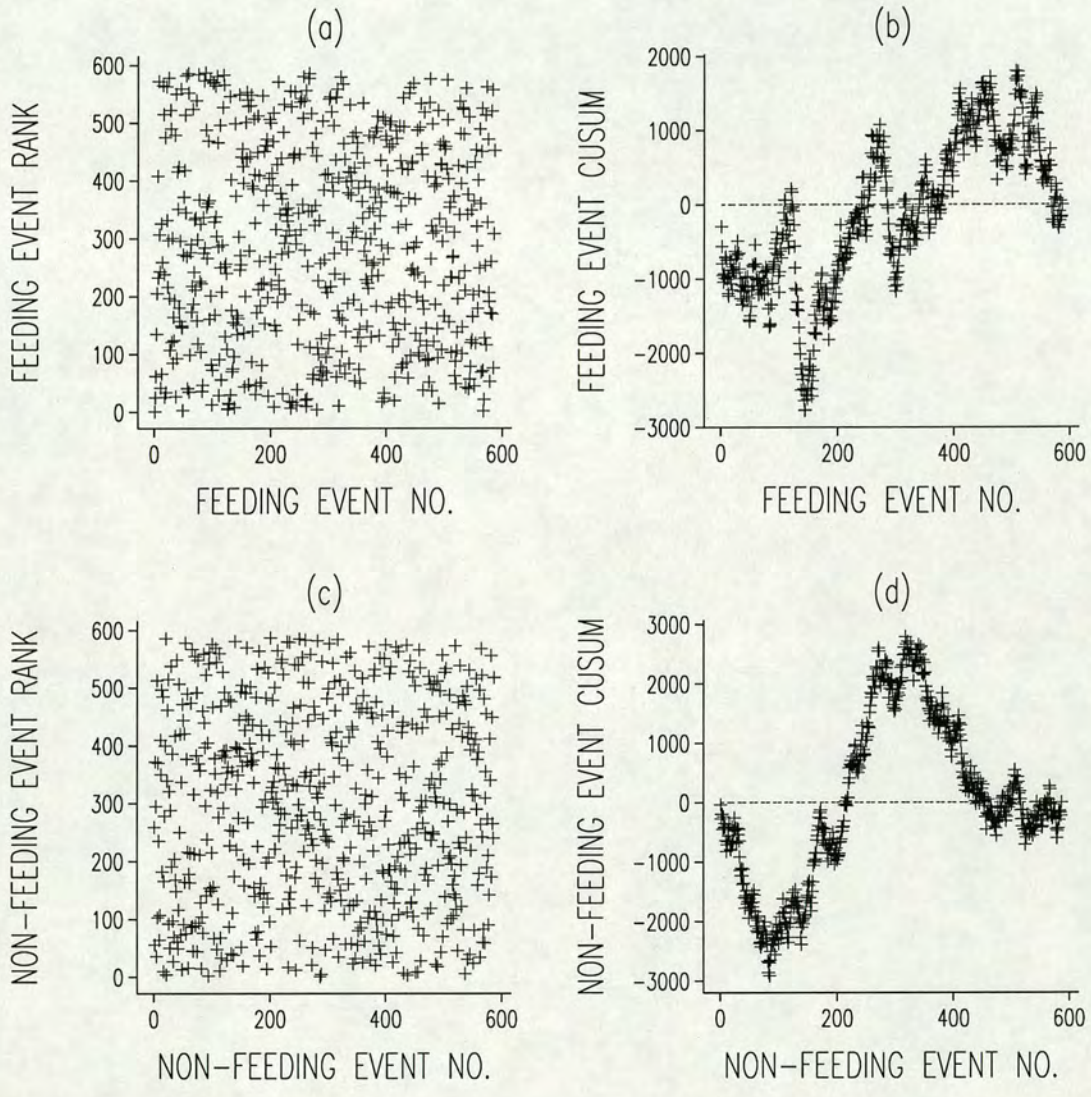

Figure 2.13: *Cow 5; (a) ranked feeding durations plotted in time order, (b) CUSUM plots for ranked feeding durations, (c) ranked non-feeding durations plotted in time order, (d) CUSUM plots for ranked non-feeding durations.*

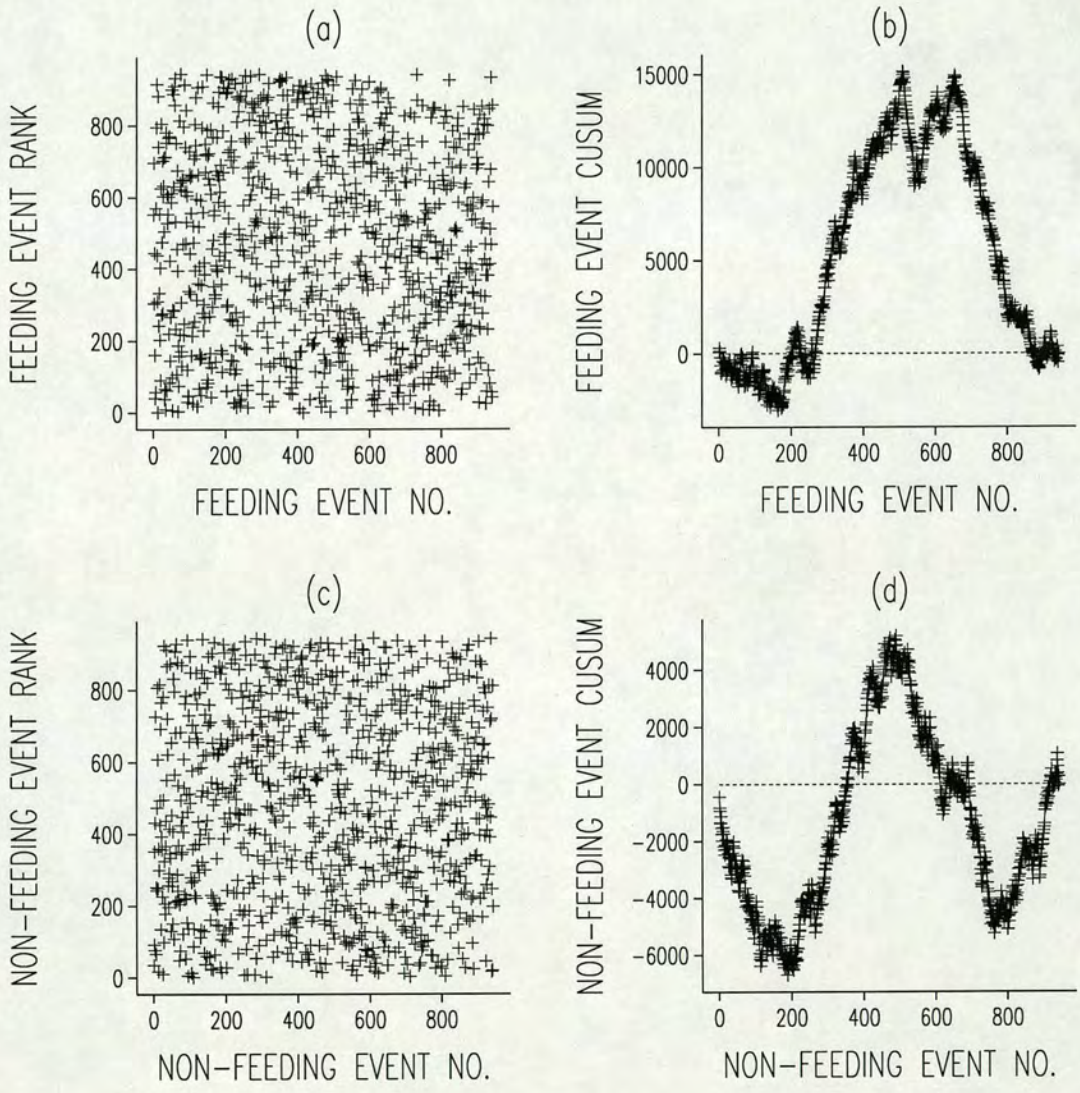

Figure 2.14: *Cow 108; (a) ranked feeding durations plotted in time order, (b) CUSUM plots for ranked feeding durations, (c) ranked non-feeding durations plotted in time order; (d) CUSUM plots for ranked non-feeding durations.*

anyway. Nevertheless this is a useful exploratory tool.

2.4 Modelling dependence

Up to now, I have ignored any dependence in the data, assuming events to be independent and looking at their marginal distributions and stationarity. I now consider to what extent the data are serially dependent. Simple pre- and post-prandial correlations are inspected, before going on to investigate whether the durations of non-feeding events are dependent on the durations of preceding non-feeding events. In the simplest case, non-feeding events can be classified as short (0) or long (1) according to the methodology in Section 2.2 and we can investigate whether the sequence of 0s and 1s formed can be considered random or whether the current type is dependent on the preceding types. Contingency table and logistic regression approaches are considered.

2.4.1 Pre- and post-prandial relationships

Figures 2.15 and 2.16 show the relationship between the duration of a feeding event and the duration of the non-feeding period preceding and following it. Because of the highly-skewed nature of the marginal distribution of non-feeding durations, these were transformed to the log scale before plotting. These relationships are commonly called the *pre-* and *post-prandial* relationships, respectively. More usually these terms would refer to meals, but here I consider the correlations in terms of individual feeder-visits. For the two cows shown, it is clear that to quote correlation coefficients for these relationships would be misleading, as any relationship is certainly not a simple linear one and in fact there do not appear to be any strong relationships at all. This seems surprising, as it might be expected that longer feeding events would be associated with longer non-feeding periods, both before and after, and although there is a hint of this in Figure 2.15, in general there is no clear relationship. Many similar findings are reported in the literature, for example Simpson (1982) looked at factors affecting meal patterns in locusts, using non-parametric tests to look for differences between their behaviour in light and dark and for trends through the day. The pre- and post-prandial relationships were generally quite poor, there being some weak evidence that meal size influenced subsequent inter-feed times, but results were largely inconclusive. Collier et al. (1999) looked at rats to see whether the strength of the relationship was related to housing conditions, believing that the

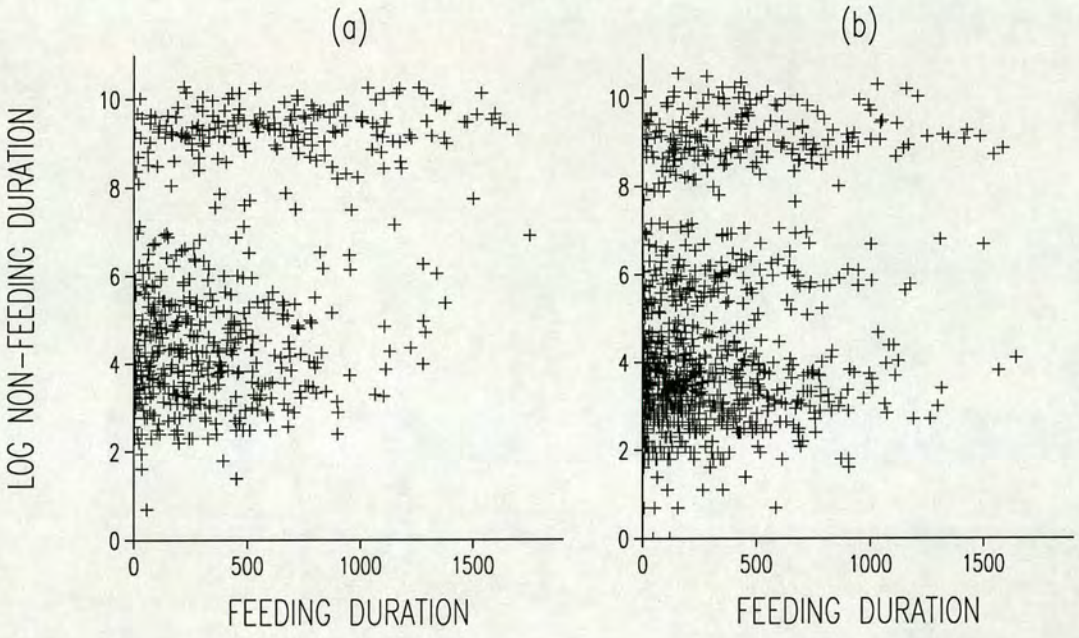

Figure 2.15: *Relationship (pre-prandial) between a feeding duration and the preceding (log) non-feeding duration; (a) Cow 5, (b) Cow 108.*

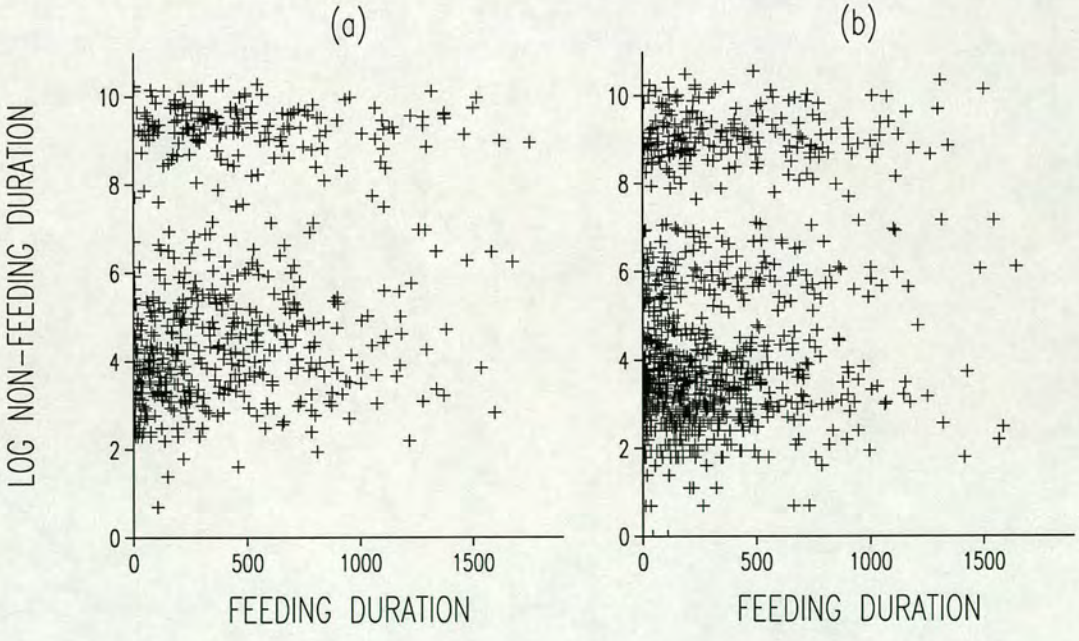

Figure 2.16: *Relationship (post-prandial) between feeding durations and the following (log) non-feeding duration; (a) Cow 5, (b) Cow 108.*

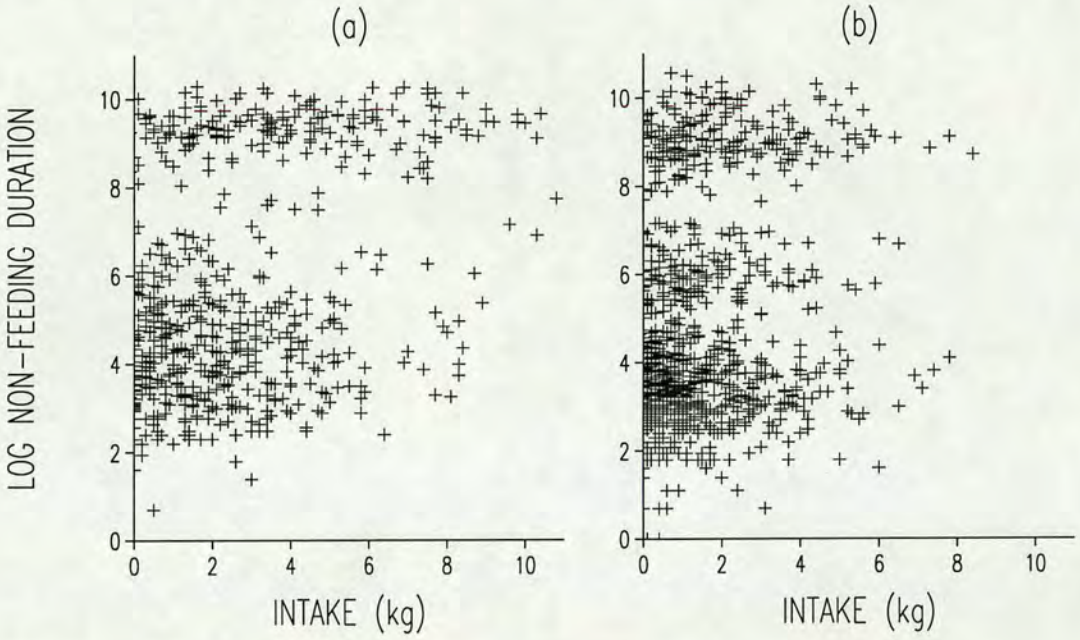

Figure 2.17: *Relationship between intake during a feeding event and the (log) non-feeding duration following it; (a) Cow 5, (b) Cow 108.*

decision to begin or end a meal depends not so much on the animal's energetic state but on the structure and economics of its habitat. However they also conclude that these correlations are unreliable. De Castro (1975) also examined rats, and found that if intake was used instead of feeding duration then stronger relationships were found. Figure 2.17 therefore shows the corresponding picture to Figure 2.16 but using intake instead of feeding duration. The two figures show essentially the same picture, suggesting that in this case, consideration of intake instead of feeding duration would offer no advantage. In neither case would it be reasonable to assume a linear model and quote correlation coefficients. In all these plots the most noticeable feature is the partitioning of the non-feeding durations into two groups, corresponding to within- and between-meal as already discussed. For Cow 108 the further partitioning into three groups can be seen.

For the same two cows, Figure 2.18 shows the relationship between the duration of a feeding event and the duration of the previous one. Again there is no evidence to support an intuitive hypothesis such as a larger feeding event following a smaller one and vice versa. Similarly, Figure 2.19 shows the relationship between adjacent non-feeding durations, both on the log scale. Again, no simple relationship is evident, but these plots display well the separation of events into short (within-meal) and long (between-meal). Therefore these plots can be partitioned into 2×2 regions corresponding to short–short, short–long, long–short and long–long

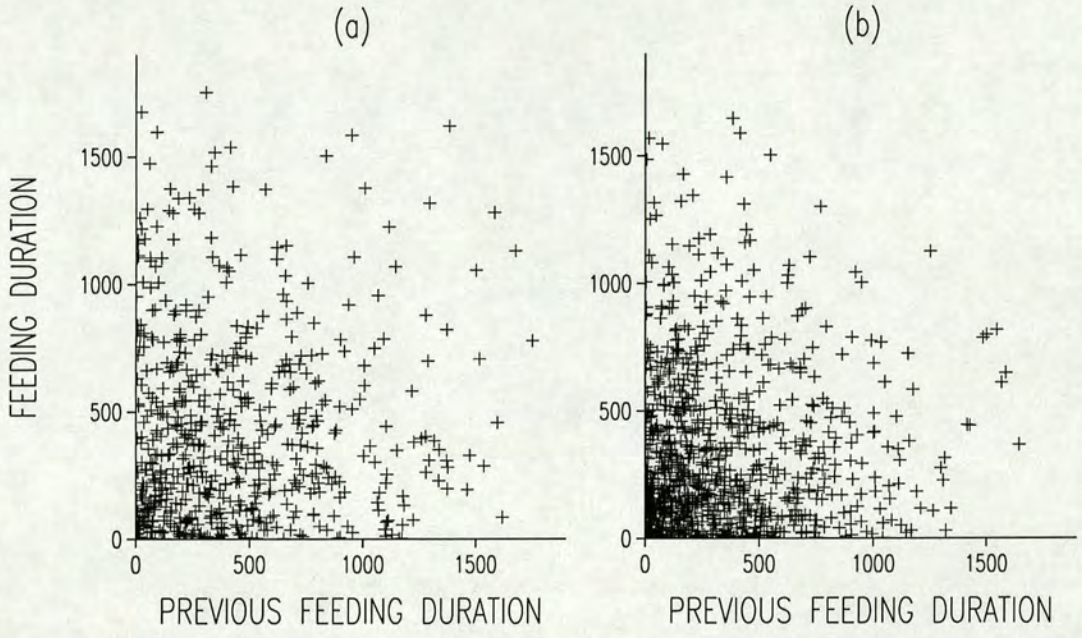

Figure 2.18: *Relationship between adjacent feeding durations; (a) Cow 5, (b) Cow 108.*

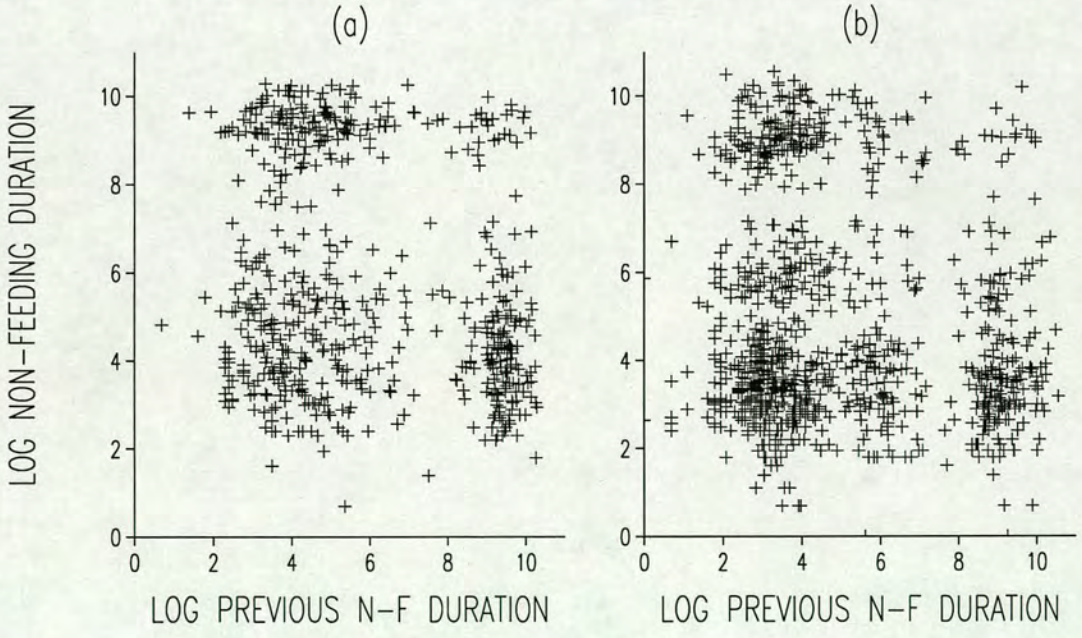

Figure 2.19: *Relationship between adjacent non-feeding durations, on the log scale; (a) Cow 5, (b) Cow 108.*

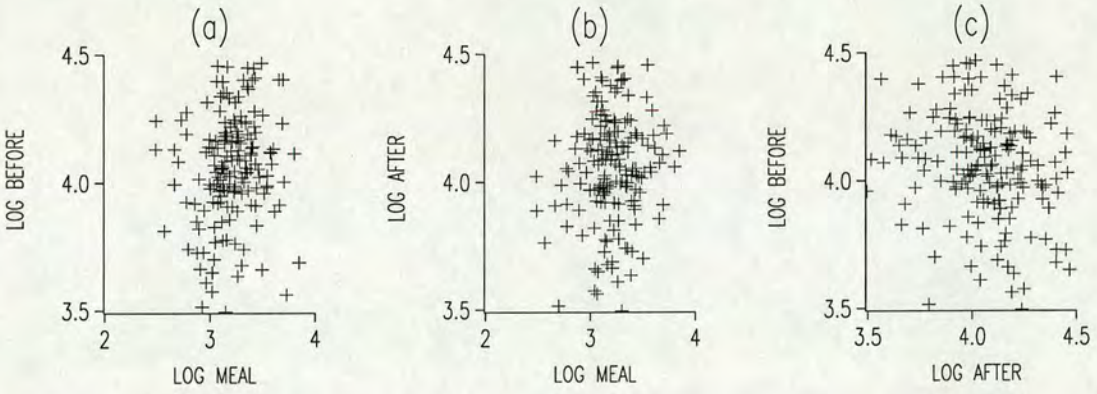

Figure 2.20: *Relationship between the duration of meals and the adjacent non-feeding durations, all on a log scale; (a) meal duration and preceding non-feeding duration, (b) meal duration and following non-feeding duration, (c) adjacent non-feeding durations.*

occurrences of events. This pattern is clear for most of the eight cows and for cows such as Cow 108, we see evidence of the three component distributions.

Finally in this section, it is of interest to briefly consider the nature of the above relationships when meal data are considered rather than visit data. Figure 2.20 shows, for Cow 108, plots of the pre- and post-prandial relationships, plus the relationship between adjacent non-feeding durations. There is no evidence here of any strong relationships between the duration of a meal and the duration of the non-feeding events before and after it, nor between the durations of adjacent non-feeding events. Similar pictures were obtained for the other high-protein cows.

2.4.2 Contingency tables

A simple way to assess the association of the current non-feeding duration with its preceding ones is to use a chi-squared test for independence. Glasbey and McGechan (1986) used such an approach, counting the number of times that two zeros occur together, the number of times 1 followed 0, etc. and so constructing the following 2×2 contingency table.

		Current	
		0	1
Previous	0	N_{00}	N_{01}
	1	N_{10}	N_{11}

For this table, the Pearson X^2-statistic can be calculated by summing $(O-E)^2/E$ over the four cells, where O is the observed count in a cell and E is the expected value under the independence assumption. The statistic is compared with a χ^2-distribution with 1 degree of freedom, lack of significance indicating no evidence against the sequence being random 0s and 1s, and that the cows' choice of whether to take a short or long gap between feeding events has nothing to do with whether the previous one was short or long. Biologically and intuitively, this would seem unlikely, and we would expect to find some evidence that the current event type is dependent on the previous one. This being the case, we can then go on to look at whether the dependence just goes back to the immediately preceding event, or whether it also depends on the one previous to that. To test this we can construct the following pair of contingency tables.

		Current	
		0	1
Previous	0 0	N_{000}	N_{001}
	1 0	N_{100}	N_{101}

		Current	
		0	1
Previous	0 1	N_{010}	N_{011}
	1 1	N_{110}	N_{111}

The first table is testing whether the current non-feeding event type depends on the event type at lag 2, given that the last non-feeding event (lag 1) was short (0), and the second does similarly given the event at lag 1 was long (1). The statistics from this pair of tables can be added together and compared with a χ^2-distribution with 2 degrees of freedom. No evidence of dependence would allow the situation to be treated as first-order Markov, i.e. only the immediately preceding event type affects the current one, whereas evidence of dependence on the event type at lag 2 would indicate a second-order Markov model.

If evidence is found of dependence at lag 2, the procedure can be continued and dependence looked for at lag 3, this time considering four contingency tables, each corresponding to conditioning the previous two non-feeding events to be 00, 01, 10 or 11. The combined statistic from these is compared against a χ^2-distribution with four degrees of freedom. Similarly we could go on and consider lag 4 (8 tables) and so on. This method has the advantage of being simple, but becomes unsatisfactory when considering higher lags, as many contingency tables need to be considered. This also leads to problems of low counts in cells, invalidating the asymptotic approximation to the χ^2-distribution.

Table 2.9 shows some results for the eight high-protein cows. There is strong evidence for all the cows that the current non-feeding event type is dependent on the previous one. From the pooled tables for dependence on the type at lag 2

Cow	*Lag 1*	*Lag 2*			*Lag 3*				
	(1)	*Single tables* (1)		*Pooled* (2)	*Single tables* (1)				*Pooled* (4)
5	25.0	5.9	0.0	5.9	9.0	0.3	2.2	0.0	11.5
41	15.6	7.2	0.3	7.5	3.5	1.7	1.9	0.7	7.9
108	23.9	5.4	0.4	5.8	17.1	2.3	0.8	0.1	20.3
169	18.2	0.9	1.8	2.8	0.0	0.7	3.4	0.2	4.3
170	26.1	9.7	0.8	10.5	7.4	0.1	1.6	0.0	9.1
182	7.6	7.3	0.9	8.2	3.2	1.2	0.4	0.5	5.4
194	34.0	20.8	0.1	20.9	5.0	1.2	4.6	0.4	11.2
221	25.4	22.3	0.3	22.6	23.2	0.6	0.5	0.0	24.3

Table 2.9: *X^2-statistics for dependency of type of non-feeding event on previous types. Numbers in brackets are the relevant degrees of freedom for those columns. The critical values for χ^2-distributions corresponding to 5% significance are* 3.8, 6.0 *and* 9.5 *for* 1, 2 *and* 4 *degrees of freedom, respectively.*

given the type at lag 1, there is evidence for five of the cows of dependence, and similarly for four cows there is evidence of dependence back to lag 3. However by this stage some of the contingency tables have very small expected frequencies in some of the cells and so some of the tests are not strictly valid. Nevertheless this is a useful way of assessing the extent of the dependence in the series.

2.4.3 Logistic regression

With the classification of non-feeding events as long or short, we can also formulate the problem as a generalised linear model, modelling the probability p of the next non-feeding duration being short (0), rather than long (1), with dependence on previous durations and on other explanatory variables. This can be done via a logistic model, with p dependent on only the immediately preceding event type indicating a first-order Markov process; dependency on the event type at lag 2, given the event type at lag 1, would be a second order Markov process, and so on. Time dependency can also be included.

A series of N non-feeding events with durations $\tau = (\tau_1, \ldots, \tau_N)$ are classified into short or long durations according to the meal criterion, i.e. $S_i = 0$ if $\tau_i < T$, the meal criterion, and $S_i = 1$ otherwise. The log-likelihood of the model can then be written as

$$\mathcal{L}(p|S) = \sum_{i=1}^{N} S_i \log p(S_{i-1}, \ldots, S_{i-r}, t) + (1 - S_i) \log(1 - p(S_{i-1}, \ldots, S_{i-r}, t))$$

where $\mathcal{L}$ is the log-likelihood for p given the N non-feeding event types $S = (S_1, \ldots, S_N)$. Here p is dependent on the immediately preceding r non-feeding event types (i.e. a Markov process of order r) and on the time of day t. There are

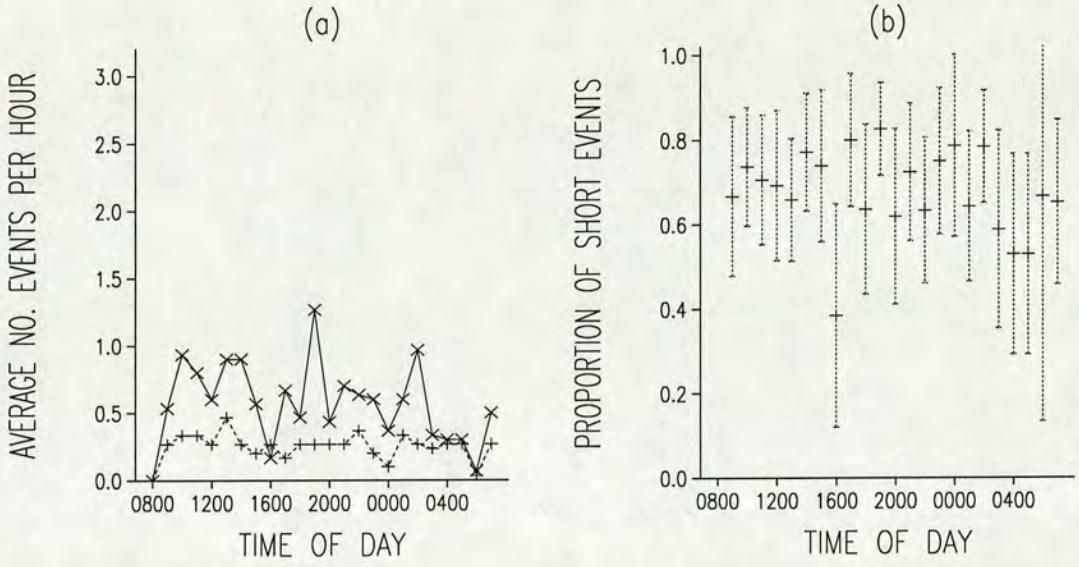

Figure 2.21: *Daily pattern of non-feeding event types for Cow 5; (a) average number of short (×) and long (+) events beginning in each hour of the day, (b) overall proportion of events beginning in that hour which are short, with confidence limits.*

clearly other potential covariates that may also be useful, for example duration and rate of eating for the immediately preceding feeding event or the type of food last eaten (for the choice cows). We constrain p to lie between 0 and 1 using a logit transformation, i.e. $\log(p/(1-p)) = f(S_{i-1}, \ldots, S_{i-r}, t)$. These models are called Markov regression models by MacDonald and Zucchini (1997, page 37).

Figures 2.21 and 2.22 illustrate, for two cows, the diurnal pattern, in terms of the numbers of short and long non-feeding events beginning during each hour of the day, and this is also expressed as a proportion of the total which begin within that hour. It would be helpful if this diurnal effect could be modelled via some sinusoidal function. However, I tried putting time of day into the model as a series of harmonics of sine and cosine terms, i.e. $2\pi jt/24$ where t is time of day in hours, and $j = 1, 2, 3, \ldots$, but no obvious pattern was found, likelihood ratio tests showing different numbers of harmonics to be significant for different cows. Therefore it was arbitrarily chosen to let the model take a different p for each hour of the day.

Table 2.10 shows results of likelihood ratio tests for models fit to the high protein cows. 0 denotes the independent model, 1 with dependence back to lag 1, $1t$ with dependence back to lag 1 and on time of day, and so on. It is clear that inclusion of terms for hour of the day and both the preceding two event types are needed in the model. For six of the eight cows there is evidence of dependence on the lag

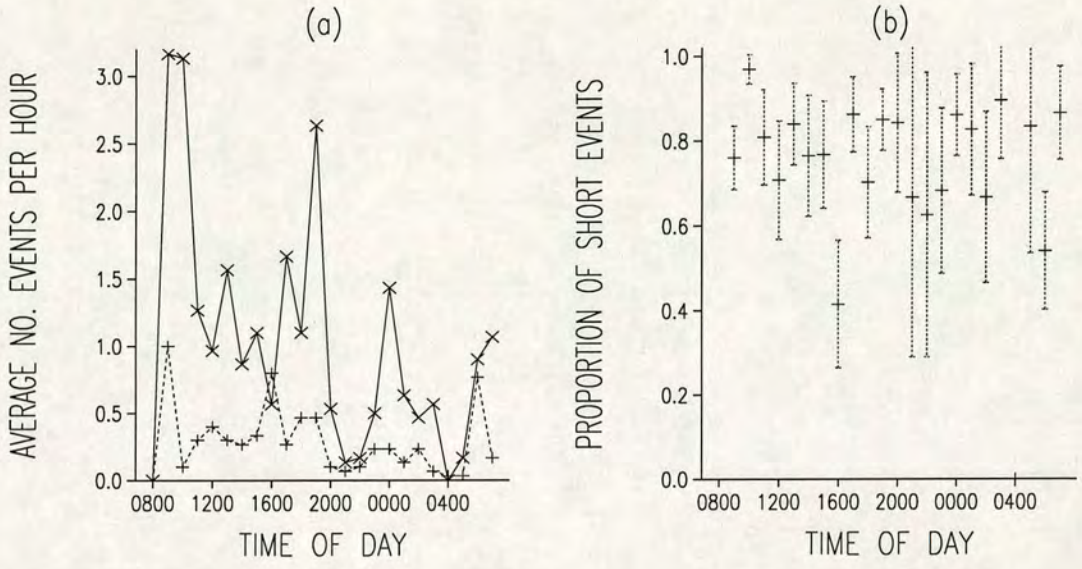

Figure 2.22: *Daily pattern of non-feeding event types for Cow 108; (a) average number of short (×) and long (+) events beginning in each hour of the day, (b) overall proportion of events beginning in that hour which are short, with confidence limits.*

Cow	0 *vs* 1 (1)	0 *vs* 0*t* (22)	1 *vs* 1*t* (22)	0*t vs* 1*t* (1)	1*t vs* 2*t* (2)	2*t vs* 3*t* (4)	3*t vs* 4*t* (8)
5	27.7	22.2	28.5	33.9	8.4	14.1	8.2
41	20.3	47.8	58.6	31.1	17.0	15.7	4.6
108	30.7	91.9	109.0	47.8	15.5	34.8	11.0
169	21.1	27.4	29.9	23.6	4.2	6.3	12.0
170	32.4	39.1	51.2	44.5	18.0	13.8	4.8
182	8.8	46.1	54.6	17.2	10.4	8.2	6.2
194	38.4	28.6	35.6	45.5	31.1	13.5	11.7
221	33.3	81.1	101.5	53.8	48.0	44.3	27.7
Critical values of χ^2-distributions							
10%	2.7	30.8	30.8	2.7	4.6	7.8	13.4
5%	3.8	33.9	33.9	3.8	6.0	9.5	15.5
1%	6.6	40.3	40.3	6.6	9.2	13.3	20.1
0.1%	10.8	48.3	48.3	10.8	13.8	18.5	26.1

Table 2.10: *X^2-statistics to compare models for p, the probability of a non-feeding event being short. The models shown range from an independent model (0) to one that includes dependence back to the event type at lag 4 and time of day (4t). Figures in brackets are degrees of freedom and critical values for the relevant χ^2-distributions are given.*

3 event type, but for only one cow is there evidence of dependence back to lag 4.

It is useful to compare these results with those given in Table 2.9, for which no allowance was made for diurnal effect. It is particularly useful to look at the first columns of Tables 2.9 and 2.10. Both are X^2-statistics for testing whether the series of 0s and 1s are independent or whether there is lag 1 dependence. The values in the former table are Pearson X^2 and are seen to be consistently lower than the values in the latter table, which are deviances from likelihood ratio tests. Both are asymptotically distributed as χ^2 with one degree of freedom, but are known to differ, especially when values are high in the tail of the distribution, as they are here (see for example Fienberg, 1980, section 3.5 and Appendix IV).

2.5 A critical timescale for feeding behaviour

A shorter form of this section was read as a paper at the *Thirty-first Meeting of the Agricultural Research Modellers' Group*, the abstract being published as Allcroft et al. (1999).

If we consider a cow's food intake over time, we see that on a short timescale, e.g. hourly or minutely, intake is very variable, some hours containing large meals and some containing no feeding at all. In contrast, for longer periods, e.g. weeks or months, intake is fairly consistent from one month to the next. So as the timescale is increased, the variability in intake decreases. The question is, does this variability simply decrease in a smooth fashion, or is there some particular length of time over which feeding there is a sudden change. This would correspond to a length of time over which a cow regulates her feeding. We hypothesise that such a timescale does exist and expect it to be of the order of a few days.

Figure 2.23 shows total daily intake for Cow 108 over the 30 days and Figure 2.24 shows histograms of the proportions of various time lengths spent eating. It shows that as the time length increases, the spread of the distribution decreases. If we take the proportion over several days, we have a distribution that can be considered normal and with small variance. Then as the time length is decreased, the distribution becomes broader. We are interested in the critical time length when the form of the distribution changes, i.e. a sudden increase in variance.

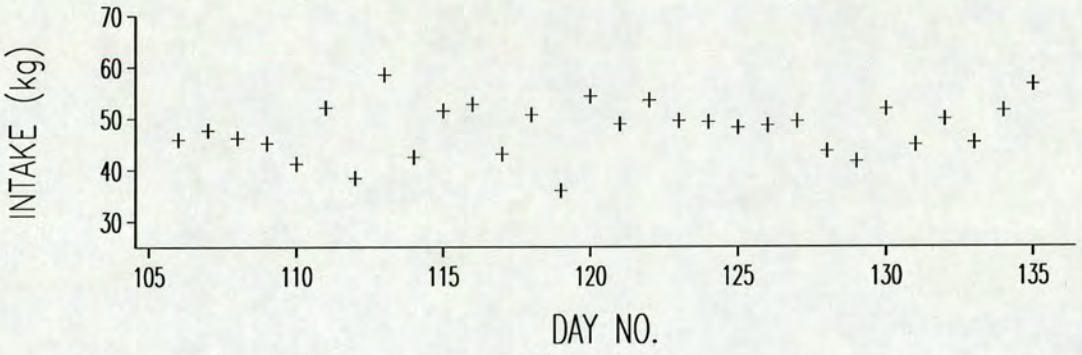

Figure 2.23: *Daily intake for Cow 108.*

2.5.1 Methodology

For each cow, let x denote a time series, with x_t the proportion of time spent eating during minute t, for $t = 1, \ldots, n$. The choice of scale of a minute is arbitrary – data were recorded in seconds, but this would result in series of prohibitive length. By aggregating to minutes we are reducing the length of the series by a factor of 60, but still retaining an acceptable amount of precision.

An example of part of one of these series is

$$x = (\ldots, 0, 0, 0.3, 1, 1, 1, 1, 1, 0.8, 0, 0, \ldots, 0, 0, 0.4, 1, 1, \ldots, 1, 1, 0.2, 0, 0, \ldots).$$

A run of 0s corresponds to a period between feeding events, a run of 1s is within a feeding event and a value between 0 and 1 is a minute during which a feeding event is begun or ended, or e.g. if both preceded and followed by a 0, a feeding event starting and ending within a single minute.

Sample autocorrelation coefficients $\hat{\rho}_l$ are easily calculated for all lags, $l = 1, 2, \ldots$ as

$$\hat{\rho}_l = \frac{1}{n} \sum_{t=1}^{n} x_t x_{t+l \bmod n}.$$

At short lags, there is positive correlation, i.e. given $x_t = 1$ it is quite likely that $x_{t+1} = 1$, and similarly given $x_t = 0$ it is quite likely that $x_{t+1} = 0$.

We can equivalently consider the variance of varying-length sums, i.e.

$$V_s = \mathrm{Var} \sum_{t=1}^{s} x_t \quad \text{for } s = 1, 2, 3, \ldots$$

e.g.

$s = 1$ gives the proportion of individual minutes spent eating,

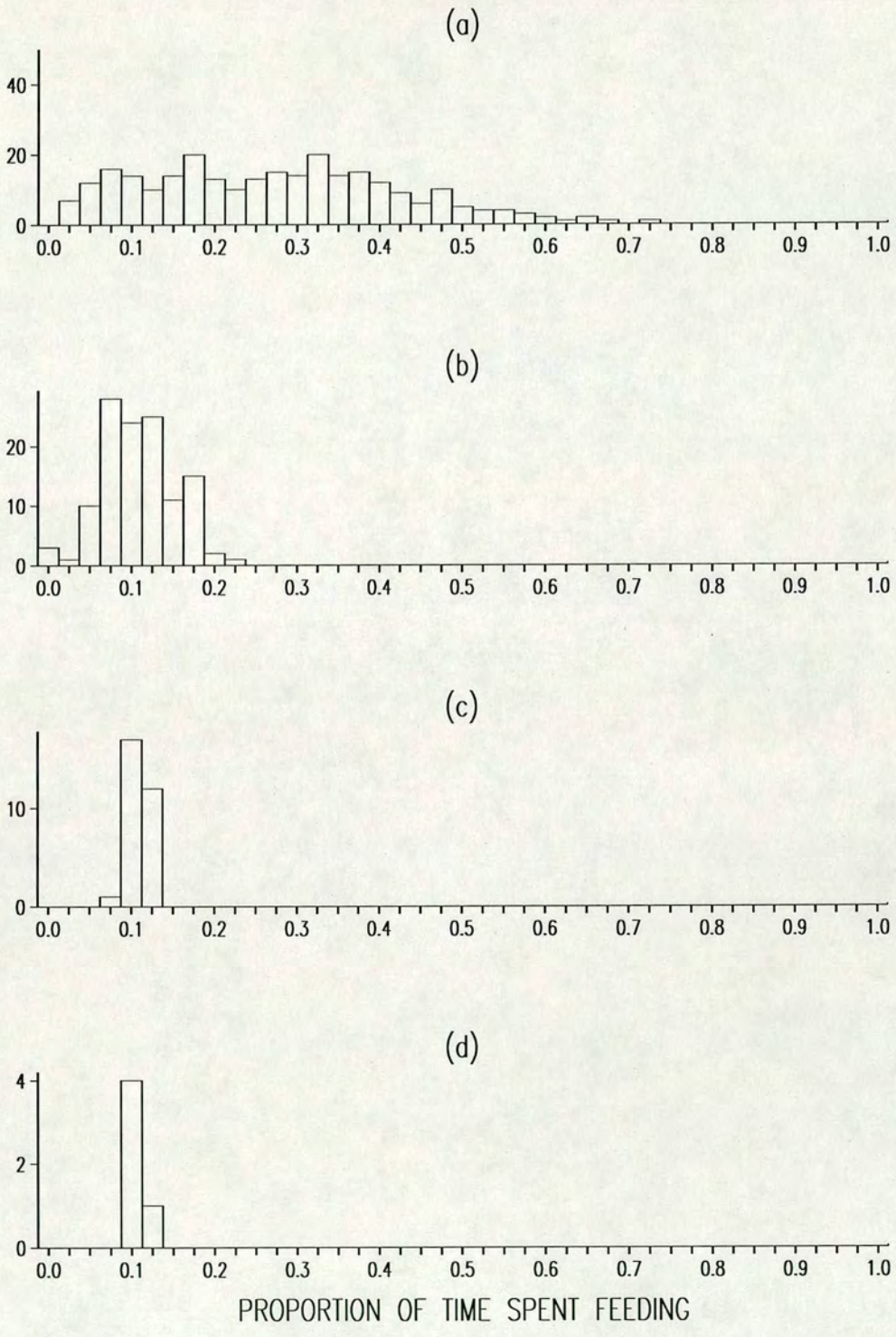

Figure 2.24: *Proportion of time spent feeding for Cow 108, averaged over different lengths of time; (a) 1 hour, (b) 6 hours, (c) 1 day, (d) 5 days.*

$s = 60$ gives the number of minutes eating per hour,
$s = 1440$ gives the number of minutes eating per day,
$s = 10080$ gives the number of minutes eating per week.

If $\text{Var}(x_t) = \sigma^2$, then the expectation of V_s is given by

$$E[V_s] = E\left[\text{Var}\sum_{t=1}^{s} x_t\right] = s\sigma^2 + 2\sigma^2\sum_{l=1}^{s-1}(s-l)\rho_l, \tag{2.1}$$

where ρ_l is the autocorrelation at lag l.

If the x_t's were independent, i.e. $\rho_l = 0, \quad \forall\ l \geq 1$, this would simply be equal to $s\sigma^2$, i.e. the variance would be proportional to s, hence V_s/s would remain constant. Therefore an investigation of how V_s/s changes with s is simply a different way of looking at how the ρ_l's differ from zero.

In the presence of non-zero autocorrelation, it is useful to consider the change in $E[V_s/s]$ as s is increased. In going from s to $s+1$ the extra correlation being added in is

$$E\left[\frac{V_{s+1}}{s+1} - \frac{V_s}{s}\right] = \frac{2\sigma^2}{s(s+1)}\sum_{l=1}^{s} l\rho_l. \tag{2.2}$$

This change is a weighted sum of the correlation coefficients up to lag s and it is the signs and relative magnitudes of these which determine how V_s/s changes with s. Consideration of V_s for $s = 1, 2, 3, \ldots$ is equivalent to looking at the set of autocorrelations ρ_l for $l = 1, 2, 3, \ldots$, but it is informative to consider both representations, as each highlights different features of the data.

2.5.2 Results

Figure 2.25 shows the autocorrelation for Cow 108 for lags up to 3 days (4320 minutes). The shape is similar for all cows, positive but decreasing correlation for the first few lags, typically up to about 40 minutes when the correlations become negative, reaching a minimum and then becoming positive again between about 100–200 minutes, after which there are oscillations about zero. For cows such as Cow 108 that exhibit a strong diurnal pattern in their feeding behaviour, there are distinct peaks in the autocorrelation at 1 day (1440 minutes) and at multiples of this thereafter. In order to see what happens to the autocorrelation in the absence of this diurnal pattern, we removed it in various ways. Parametrically we tried estimating x_t by a series of sine and cosine terms

$$\hat{x}_t = \bar{x} + \sum_{j=1}^{J}\left(a_j \sin\frac{2\pi jt}{1440} + b_j \cos\frac{2\pi jt}{1440}\right),$$

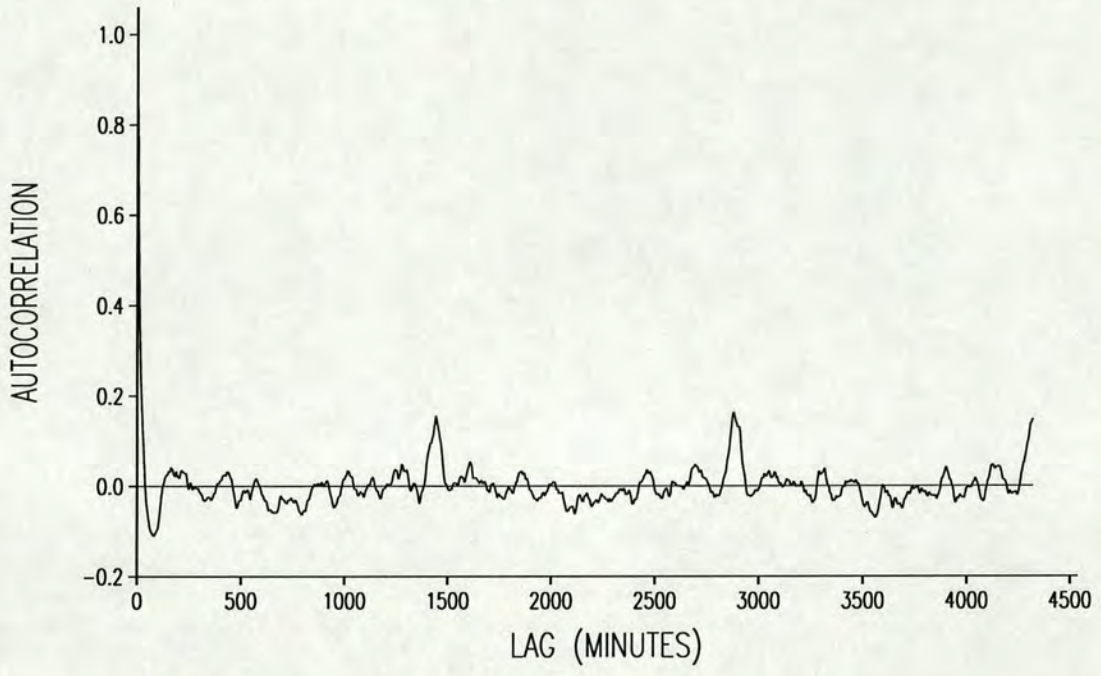

Figure 2.25: *Estimated autocorrelation for Cow 108.*

where t is in minutes and J was arbitrarily chosen to be either 12 or 24, both giving similar results. As a non-parametric approach, minutely means were calculated over the 30 days and a moving average was used to smooth (see also Section 3.7.1.2). We can then consider the autocorrelation structure of the series $(x_t - \hat{x}_t)$, shown in Figure 2.26. The overall shape is similar to before, but the peaks at multiples of whole days have now been removed.

Going on to the other approach, Figure 2.27 shows the equivalent information to Figure 2.25 in terms of variances. Again the picture is clearly distorted by the diurnal pattern, therefore Figure 2.28 shows the picture after diurnal pattern has been removed. Inspection of such plots for all eight high-protein cows showed them all to be of the same shape. V_s/s increases with s up to a maximum at around $s = 1$ hour and then decreases. Actual positions of the maxima for the 8 cows are shown in Table 2.11, both before and after adjustment for diurnal effect. The positions of the maxima are seen to be fairly consistent over cows. All decrease after adjustment for diurnal effect, as expected, since removing the diurnal effect is removing some of the correlation. After the maximum, the graphs decrease up to a time $s = 1$ day and then remain fairly level. Cows with a strong diurnal pattern show troughs at whole days for the unadjusted data. Allowance for diurnal effect removes this feature and the graphs are all fairly level after a

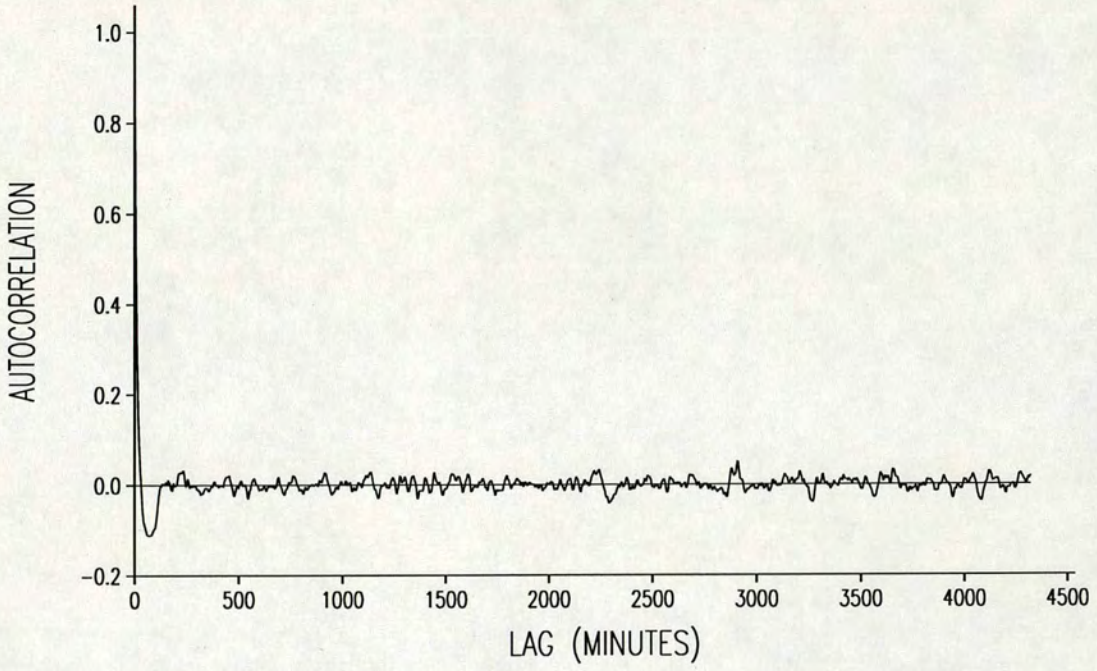

Figure 2.26: *Autocorrelation for Cow 108 after adjustment for diurnal trend using a moving average of length one hour.*

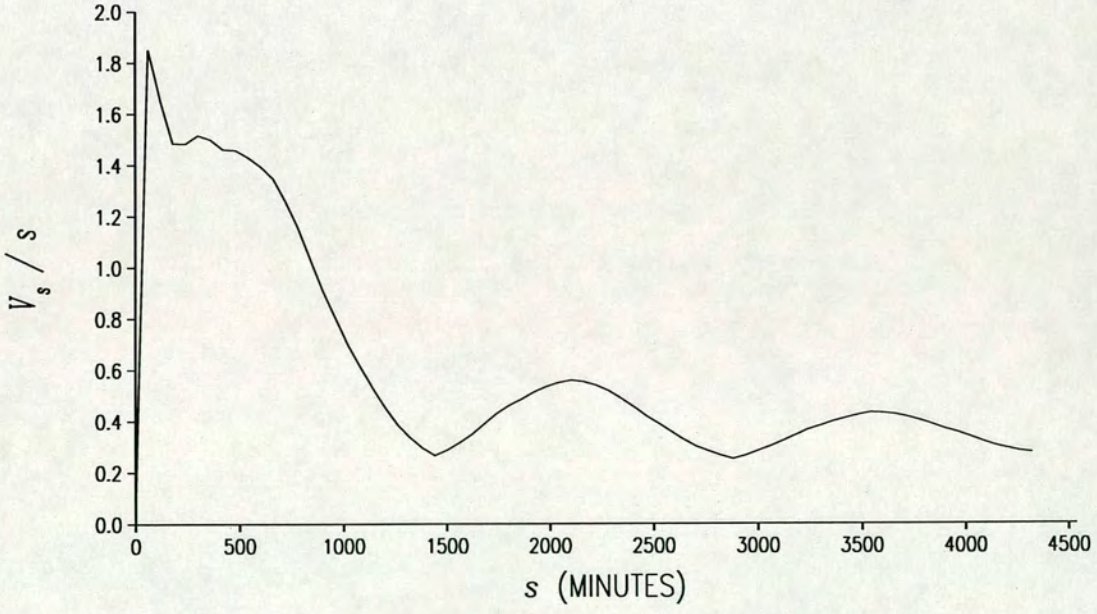

Figure 2.27: *V_s/s plotted against s for Cow 108, with no adjustment for diurnal trend.*

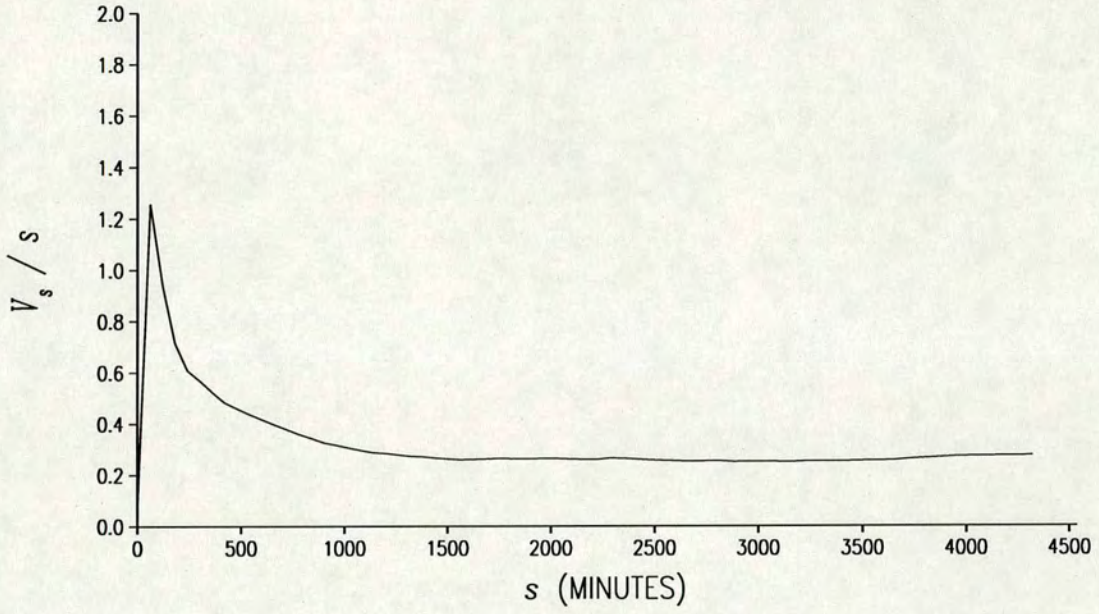

Figure 2.28: *V_s/s plotted against s for Cow 108 after adjustment for diurnal trend using a moving average of length one hour.*

Cow	s_{max}	
	Unadjusted	*Adjusted*
5	63	58
41	80	72
108	69	55
169	66	56
170	77	69
182	99	84
194	58	44
221	71	63
Average	73	63

Table 2.11: *Position of maxima s_{max} (in minutes) of V_s/s for the eight high-protein cows, with and without adjustment for diurnal effect.*

time of 1 day, as in Figure 2.28.

2.5.3 Interpretation of plots

Biologically we can try and relate the autocorrelation to meal durations and inter-meal durations. For example positive autocorrelation up to about 40 minutes ties in with whole meals typically being of around this length, and similarly a trough at 80 minutes means that given the cow is eating now, she is unlikely to still be eating (or have embarked on another meal) 80 minutes later. For the cows which exhibit little daily consistency in their meal times, e.g. Cow 169, we just appear to have irregular oscillations at higher lags; for those which have more daily structure in their feeding times, e.g. Cows 108 and 221, we see more regular oscillations at later lags, as here it is easier to predict when the next meal is likely to be. Note that the size of these oscillations are reduced after diurnal effect has been removed.

The plots of V_s/s are less obvious to interpret biologically. The increase up to 1 hour indicates that in this part of the graphs, the weighted sum of correlations (2.2) being added into (2.1) as we increase s is more positive than negative. For small s this is clear as all the ρ_l are positive. For s between 40 and 60 minutes, even though the autocorrelation is generally negative above lag 40, the positive correlation from the smaller lags is outweighing the contribution from the negative correlation at the higher lags. After the peak at 1 hour the opposite is true and the higher lag negative correlations dominate and hence the graphs decrease.

A high value of V_s/s therefore indicates that the sum in (2.1) is dominated by positive correlations. The peak at 1 hour is telling us that sums over whole hour periods are more variable than any other length sums. This makes sense if we relate 1 hour to about half the time length between starts of meals — then by summing over whole hours we will have quite a few sums which contain no eating and others which contain whole meals. Thinking about sums of greater length than this, these are more likely to contain substantial periods of both feeding and non-feeding, leading to relatively lower variation; this gets more true as the length of the sum is increased. Hence we see the relative variation decreasing as we move from an hour to a day. In this part of the graph the longer-term negative correlation is having more impact on the variance than the short-term positive components, i.e. the intake is being regulated. After 1 day, V_s/s remaining roughly constant indicates that there is no more regulation over several days than over a single day.

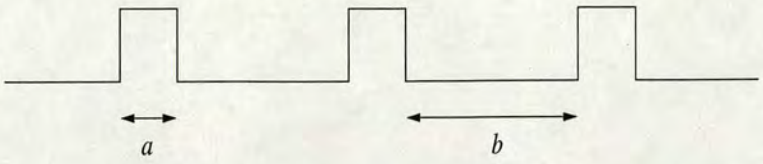

Figure 2.29: *An artificial series with regular periods of feeding of duration a minutes, separated by regular non-feeding periods of b minutes.*

Therefore we must conjecture that we have a critical timescale of a single day. Use of this technique has produced no evidence of feedback over longer times than this.

2.5.4 Comparison with an artificial series

To further investigate the peak around 1 hour, an artificial series was created as illustrated in Figure 2.29, in which the cow alternates between eating for a minutes and not eating for b minutes. V_s/s for this series can be derived as a piecewise analytic function, the position of the maximum of which depends on the relative sizes of a and b.

The position of the maximum is given by

$$s_{max} = \begin{cases} \sqrt{\dfrac{(b^2-1)(a+b)}{3b}} & \text{for} \quad b < a/2 \\ \dfrac{3ab}{2(a+b)} & a/2 \leq b \leq 2a \\ \sqrt{\dfrac{(a^2-1)(a+b)}{3a}} & b > 2a \end{cases} \tag{2.3}$$

These maxima are derived from explicit expressions for V_s/s. In the case of $b > 2a$, which is relevant to the data we consider, we have

$$\text{Var}\left(\sum_{i=1}^{s} x_i\right) = \begin{cases} \dfrac{s}{3(a+b)^2}\Big(3sab - (s^2-1)(a+b)\Big) & \text{for} \quad s \leq a \\ \dfrac{a}{3(a+b)^2}\Big(3sa(b+a-s) - (a^2-1)(a+b)\Big) & a \leq s \leq b \\ \dfrac{a}{3(a+b)^2}\Big(3sa(b+a-s) - (a^2-1)(a+b) & \\ +\dfrac{1}{a}(a+b)(s-b-1)(s-b)(s-b+11)\Big) & s \geq b \end{cases}$$

Figure 2.30 shows the curve for typical values of $a = 40$ minutes feeding followed by a non-feeding period of $b = 220$ minutes. Solving (2.3) for $a = 40, b = 220$

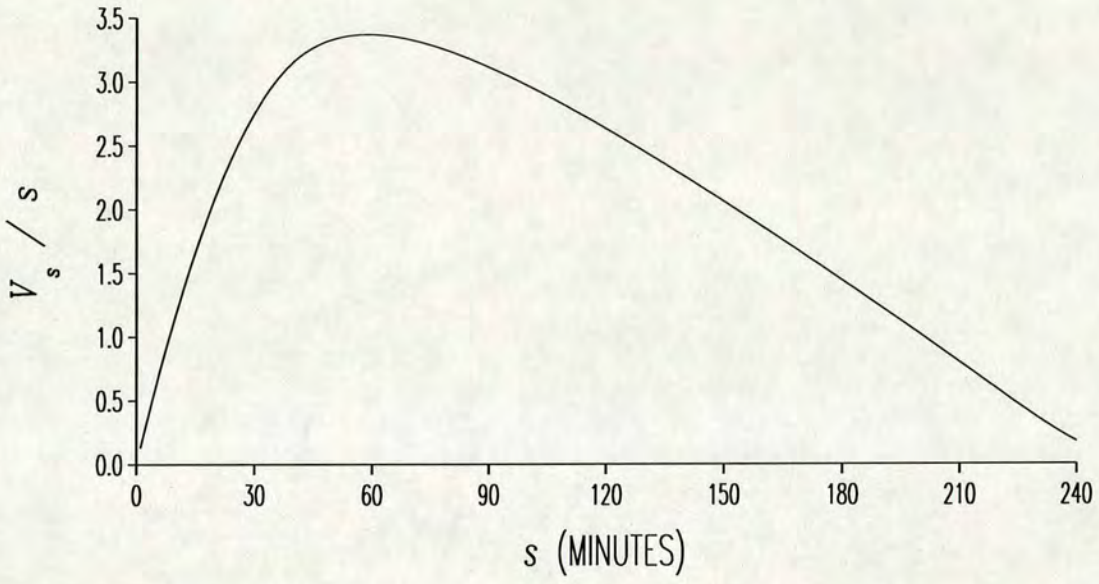

Figure 2.30: *V_s/s plotted against s for an artificial series with regular meal durations of 40 minutes and inter-feed durations of 3 hours 40 minutes.*

gives $s_{max} = 59$ minutes, i.e. the peak occurs at around 1 hour. This is obviously only a crude approximation of the real situation, but it is reassuring that we have a peak at a similar position as that observed from the sets of data from the eight high-protein cows. Hence such a peak can be considered to be an artifact of this sort of series.

Figure 2.31 shows the position of the maximum for other combinations of values of a and b. The line labelled (3) corresponds to a maximum at 60 minutes and so any combination of meal duration a and inter-meal duration b that lie on this line also produce a peak at around 1 hour.

2.5.5 Other approaches

The above work involves fixing a length of time, s, and looking at the variance of the proportion of time spent eating over this length of time, over the whole series. An alternative approach would be to consider different quantities within each segment, e.g. given the total amount eaten over the whole segment, we could look at the maximum deviation of the cumulative amount eaten from the expected amount. We can create all windows of length s, find the maximum deviation within each window and then average over all the windows and compare with the expected value. The average deviation and the average squared deviation could

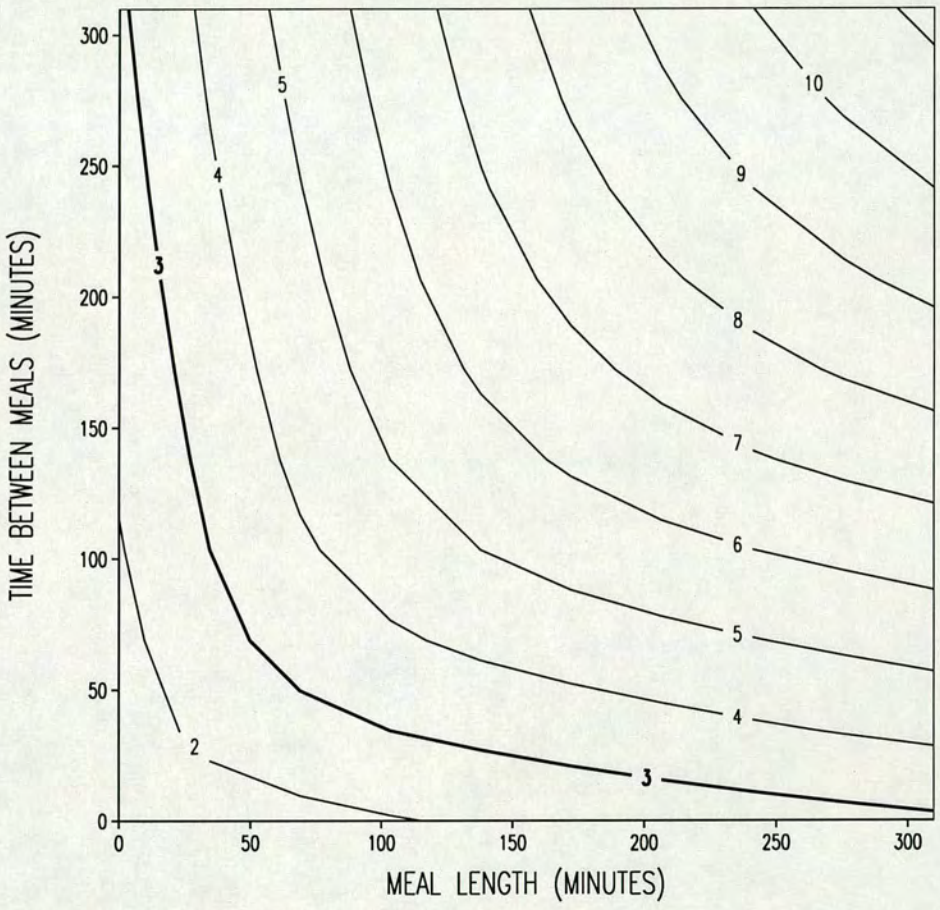

Figure 2.31: *Contour plot to show position of maximum for relative values of a and b. Key is (1)=20 minutes; (2)=40 minutes; (3)=60 minutes; up to (11)=220 minutes. (3) is highlighted as being close to what was observed from the data.*

be used similarly.

Another idea would be to consider these statistics in relation to fixing an amount of food and considering the variation in time taken to eat the fixed amount of food. Yet another approach would be to look at statistics within these windows, as opposed to between them. However consideration of these alternative approaches appeared to produce no further interesting results and so details are not presented here.

2.6 Summary

In this chapter, various approaches to modelling have been considered, all of which have been suggested by the empirical form of the data. Firstly, the high positive correlation between intake and feeding duration means that instead of modelling intake directly, which might be the biologically obvious thing to do, intake can be assumed to have a constant rate and feeding durations can be modelled instead. For the marginal distributions of feeding durations, exponential distributions were found to be a good description, and for non-feeding durations, mixtures of log-normal distributions were seen to be a better fit than the mixtures of exponential distributions common in the literature. This also has a better biological basis and allows meal criteria to be well-determined. Overall trend in the data series was addressed by inspection of CUSUMs, finding that in some of the series there was evidence of non-stationarity, however this may be due to serial dependence rather than overall non-stationarity. To assess the extent of dependence in the data, plots were examined to look for relationships between feeding and non-feeding durations. Pre- and post-prandial correlations were found to be unreliable, and dependence of non-feeding events, after classification as short or long, was investigated by Pearson chi-squared tests and logistic modelling, the latter having scope for extension to inclusion of covariates other than time. Finally, the existence of a critical timescale for cow feeding behaviour was investigated by consideration of the autocorrelation and variance properties of the series. Some interesting results were found, but no evidence of a critical timescale longer than a day.

Chapter 3

Latent Gaussian model

I consider a model for which the observed binary data are considered to have arisen from the thresholding of a latent Gaussian variable. The biological motivation is discussed in Section 3.1. In Section 3.2, for convenience, I present the notation used for the specific classes of ARMA process that are used later on. The estimation of the autocorrelation of the observed and latent processes are discussed in Section 3.3 and the relationship between then is highlighted. Section 3.4 then describes some computationally-fast methods for model estimation, including methods based on least squares, pairwise likelihood and the spectral representation, the motivation for the latter being dealt with fully in Chapter 4. In Section 3.5, an efficient but computationally-intensive method of parameter estimation using Markov chain Monte-Carlo (MCMC) is considered. Section 3.6 describes a simulation study which shows that for a range of ARMA processes, the spectral method can be more efficient than variants of least squares and much faster than MCMC. In Section 3.7, I return to the data and illustrate the fitting of an ARMA(2,1) model. A summary of much of the work in this and the next chapter is contained within Allcroft and Glasbey (2000, 2001).

3.1 Motivation

In its simplest form, feeding can be considered to take the form of binary time series, the two possible states being feeding and non-feeding. Previously, because of the useful analytic properties of Gaussian variables, a transformation has been applied to similar types of data in order to achieve normality. For example, Glasbey and Nevison (1997) apply a monotonic transformation to rainfall data to achieve marginal normality. This defines a *latent Gaussian variable*, for which zero rainfall corresponds to censored values below a threshold, and when raining,

the actual value of the variable is known. For the feeding data, the idea is to create an artificial normally distributed variable from the data for which periods of feeding correspond to the latent variable exceeding some threshold. Unlike the rainfall data, we only know whether we are above or below the threshold, we never know the actual value of the variable.

The biological justification for this type of model is to think of the latent variable as corresponding to some physiological or neurological states or chemical levels within the animal which affect its motivation to feed. The level of this variable changes continuously, but when it crosses the threshold, the animal is motivated to resume or stop feeding. For categorical data, continuous latent variables appear to offer a flexible approach to modelling, allowing the inclusion of diurnal cycles, covariates and multivariate dependencies between animals to be built into the model.

Figure 3.1 shows simulations of such latent and thresholded processes, in this case an ARMA(2,1) model with unit variance. Note that a deterministic link between the latent and categorical variables has been used, as for example suggested by Cox and Snell (1989, pages 101–102) and discussed extensively by Kedem (1980). An alternative would have been a stochastic link such as a logistic response, see for example Keenan (1982).

The consideration of this type of model for the cow feeding data motivates an investigation into computationally-fast methods for the estimation of parameters in ARMA processes for which the data are missing or censored. Likelihood expressions are complicated when data are censored, so our approach is to first estimate the autocorrelation of either the observed binary series or the unobserved Gaussian series, and then estimate the ARMA parameters by matching the sample autocorrelation coefficients with their expected values. ARMA processes are short-term memory processes and so for all methods of estimation, the effect of using the sample correlation at lower lags only is investigated. The general problem of missing data remains to be investigated — here I consider only the special case where the data are thresholded, as for the feeding data.

3.2 Notation

I outline definitions and notation for the special cases of the general pth-order autoregressive qth-order moving average ARMA(p, q) process that are used later in the chapter. AR(1), MA(1) and ARMA(1,1) processes are considered in detail

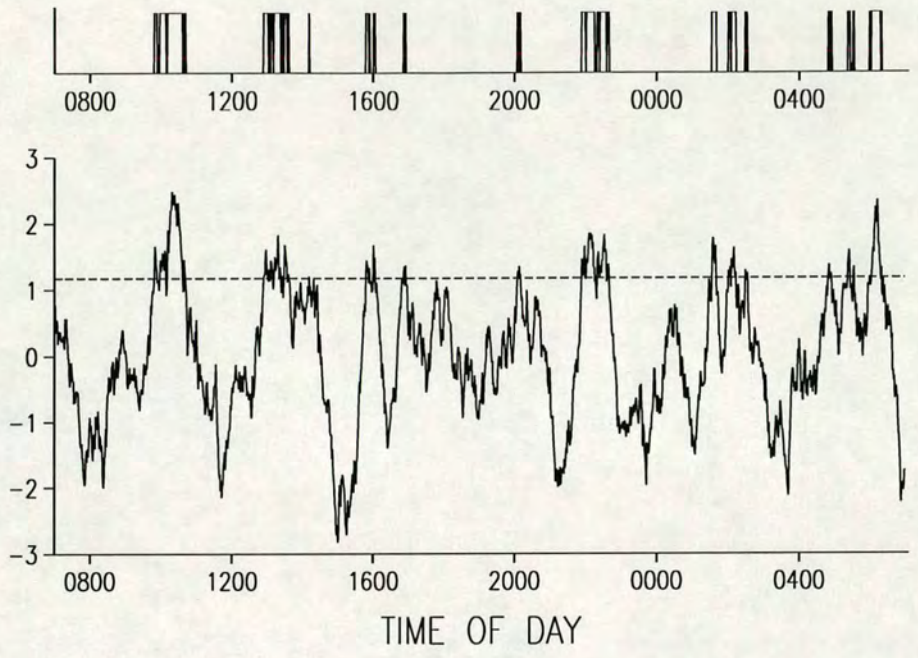

Figure 3.1: *A simulated ARMA(2,1) series, thresholded at a level to match the observed feeding rate of Cow 108. Corresponding feeding events are shown above.*

in the simulation study and ARMA(2,1) is a suitable candidate for modelling the feeding data. Similar notation and expressions and more details of the processes can be found in e.g. Box et al. (1994).

The general ARMA(p, q) process is written as

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q},$$

where y_t are observations at equidistant, discrete timepoints; ϵ_t are independently distributed normal observations, with zero mean and variance σ^2; $\phi_j, j = 1, \ldots, p$, are the set of autoregressive parameters and $\theta_j, j = 1, \ldots, q$, are the set of moving average parameters.

For the process to be stationary, the autocovariance and autocorrelation matrices are required to be positive-definite. The conditions are satisfied if the roots of the characteristic equation

$$1 - \phi_1 B - \ldots - \phi_p B^p = 0$$

lie outside the unit circle. Here, B is the backshift operator. For the process to be invertible, a condition which assures that a given model has a unique representation, the roots of

$$1 - \theta_1 B - \ldots - \theta_q B^q = 0$$

must lie outside the unit circle. For more details, see for example Box et al. (1994, sections 3.1 and 3.4).

3.2.1 AR(1) process

The first-order autoregressive process, written AR(1), is defined by

$$y_t = \phi y_{t-1} + \epsilon_t \quad \text{where} \quad \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2).$$

To satisfy stationarity conditions we need $|\phi| < 1$. The variance of y_t, σ_y^2, and the autocorrelation at lag l, ρ_l, are given by

$$\begin{aligned} \sigma_y^2 &= \sigma^2 \frac{1}{1-\phi^2}, \\ \rho_l &= \phi^{|l|} \qquad \text{for } l = 0, 1, 2, \ldots \end{aligned}$$

3.2.2 MA(1) process

The first-order moving average process, MA(1), is given by

$$y_t = \epsilon_t - \theta \epsilon_{t-1} \quad \text{where} \quad \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2).$$

To satisfy invertibility conditions we need $|\theta| < 1$. The variance and the autocorrelation at lag l are respectively given by

$$\begin{aligned} \sigma_y^2 &= \sigma^2(1+\theta^2), \\ \rho_l &= \begin{cases} 1 & \text{for } l = 0, \\ \dfrac{-\theta}{1+\theta^2} & l = 1, \\ 0 & l \geq 2. \end{cases} \end{aligned}$$

3.2.3 ARMA(1,1) process

The first-order autoregressive–first-order moving average process, ARMA(1,1), is given by

$$y_t = \phi y_{t-1} + \epsilon_t - \theta \epsilon_{t-1} \quad \text{where} \quad \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2).$$

Stationarity and invertibility require that $|\phi| < 1$ and $|\theta| < 1$. The variance and autocorrelation structure are given by

$$\sigma_y^2 = \sigma^2 \frac{1 + \theta^2 - 2\phi\theta}{1 - \phi^2},$$

$$\rho_l = \begin{cases} 1 & \text{for } l = 0, \\ \dfrac{(1-\phi\theta)(\phi-\theta)}{1+\theta^2-2\phi\theta} & l = 1, \\ \phi\rho_{l-1} & l \geq 2. \end{cases}$$

3.2.4 ARMA(2,1) process

The second-order autoregressive–first-order moving average process, ARMA(2,1), is given by

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t - \theta\epsilon_{t-1} \quad \text{where} \quad \epsilon_t \sim \text{i.i.d. } N(0, \sigma^2).$$

Stationarity requires $\phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1$ and $|\phi_2| < 1$, and invertibility requires $|\theta| < 1$. The variance and autocorrelation are given by

$$\sigma_y^2 = \sigma^2 \frac{(1-\phi_2)(1+\theta^2) - 2\phi_1\theta}{(1-\phi_2)(1-{\phi_1}^2-{\phi_2}^2) - 2{\phi_1}^2\phi_2},$$

$$\rho_l = \begin{cases} 1 & \text{for } l = 0, \\ \dfrac{(1-\phi_1\theta)(\phi_1-\theta) + {\phi_2}^2\theta}{(1-\phi_2)(1+\theta^2) - 2\phi_1\theta} & l = 1, \\ \phi_1\rho_{l-1} + \phi_2\rho_{l-2} & l \geq 2. \end{cases}$$

Although this recurrence relation only holds for $l \geq 2$, the autocorrelation function can be written as a mixture of two exponential functions for all $l \geq 0$. In fact for a general ARMA(p, q) process with $p > q$, the autocorrelation function can be written as a mixture of p exponential functions for all lags $l \geq 0$. For a fuller discussion of this see Section 3.7.2.1.

Note that for all the above processes, the formulae for variance and autocovariance, and hence autocorrelation, can be worked out directly from the definition of the process, using the standard formulae $\text{Var}(aX) = a^2\,\text{Var}(X)$ and $\text{Cov}(aX, bY) = ab\,\text{Cov}(X, Y)$, and the facts that $\text{Var}(y_{t-j}) = \text{Var}(y_t)$ for all j, $\text{Cov}(\epsilon_{t-j}, \epsilon_t) = 0$ for all j, $\text{Cov}(y_t, \epsilon_t) = \sigma^2$ and $\text{Cov}(y_t, \epsilon_{t+j}) = 0$ for all $j > 0$.

3.3 Estimation of Autocorrelation

Given the observed binary series we are assuming an underlying Gaussian process for which feeding is linked deterministically with the latent variable being above a given threshold. It is clearly not possible to determine the complete realisation of the Gaussian series from the observed binary series; the best we can do is to

estimate the autocorrelation of the Gaussian series, and in fact this is all we need, because this completely characterises the Gaussian process. It is important to note that there is a direct one-to-one correspondence between the autocorrelation of the observed binary series and the latent Gaussian series and we consider the form of this relationship. The issue of how to deal with non-stationary series is also addressed.

3.3.1 Circularity

Throughout this work, the autocorrelation structure of the series is considered to be of a circular nature, i.e. $\rho_l = \rho_{n-l}$, so the full set of expected autocorrelation coefficients for a series of length n can be written

$$(\rho_0, \rho_1, \rho_2, \ldots, \rho_{n-1}) = (\rho_0, \rho_1, \rho_2, \ldots, \rho_{\frac{n}{2}-1}, \rho_{\frac{n}{2}}, \rho_{\frac{n}{2}-1}, \ldots, \rho_2, \rho_1). \tag{3.1}$$

This is equivalent to assuming that the series repeats itself after it has finished. This might seem an artificial assumption to make, but it simplifies the mathematics considerably and has negligible effect on numerical results, especially for long series and at short lags, which will be seen to be the most important.

3.3.2 Tetrachoric and binary correlation coefficients

The autocorrelation coefficients of the latent Gaussian process can be estimated from the binary series by *tetrachoric correlation coefficients*, described in Johnson and Kotz (1972, pages 117–118). Calculations involve considering each lag in turn, $l = 1, 2, 3, \ldots$ and maximising the bivariate likelihood of the observed counts.

Let the binary series be x_t, $t = 0, \ldots, n-1$, and the underlying Gaussian series be y_t, $t = 0, \ldots, n-1$. For lag l, assuming the circular correlation structure, we consider all n pairs in the binary series at a time lag l apart, i.e. (x_t, x_{t+l}), and form counts of the number of times that each of the pairs $(0,0), (0,1), (1,0), (1,1)$ occur, so forming the following tetrachoric table.

		x_{t+l}		
		0	1	
x_t	0	a	b	$a+b$
	1	b	d	$b+d$
		$a+b$	$b+d$	n

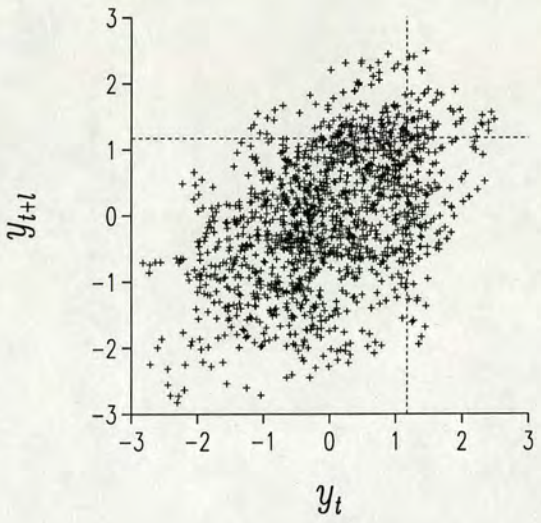

Figure 3.2: *For a fixed lag l, pairs of points (y_t, y_{t+l}) of the underlying Gaussian series are plotted. Dotted lines are at the estimated threshold $\hat{T}$, hence the numbers of points in the four quadrants are the counts in the tetrachoric table; starting in the lower left quadrant and proceeding clockwise, the counts are a, b, d, c, respectively.*

The situation is displayed graphically in Figure 3.2. The threshold of the underlying Gaussian series, $\hat{T}$, is estimated as the deviate of the standard normal distribution for which $\Phi(\hat{T})$ is equal to the overall probability of non-feeding. Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. So we estimate

$$\hat{T} = \Phi^{-1}\left(\frac{a+b}{n}\right).$$

The tetrachoric correlation coefficient at lag l, $\hat{\rho}_l^{(G)}$, is that which solves

$$\frac{d}{n} = \int_{y_t=\hat{T}}^{\infty} \int_{y_{t+l}=\hat{T}}^{\infty} f(y_t, y_{t+l}; \hat{\rho}_l^{(G)})\, dy_t\, dy_{t+l} \tag{3.2}$$

where $f(y_i, y_j; r)$ is the bivariate standard normal probability density function with correlation r, i.e.

$$f(y_i, y_j; r) = \frac{1}{2\pi(1-r^2)^{1/2}} \exp\left[-\frac{1}{2(1-r^2)}(y_i^2 - 2ry_iy_j + y_j^2)\right].$$

If we now consider expected values, we can derive a functional relationship between the binary and Gaussian autocorrelation, $\rho_l^{(B)}$ and $\rho_l^{(G)}$ respectively,

$$\begin{aligned} \rho_l^{(B)} = \frac{\gamma_l^{(B)}}{\gamma_0^{(B)}} &= \frac{E(X_tX_{t+l}) - E(X_t)^2}{E(X_t^2) - E(X_t)^2} \\ &= \frac{\mathrm{P}(X_t = 1, X_{t+l} = 1) - \mathrm{P}(X_t = 1)^2}{\mathrm{P}(X_t = 1) - \mathrm{P}(X_t = 1)^2} \end{aligned} \tag{3.3}$$

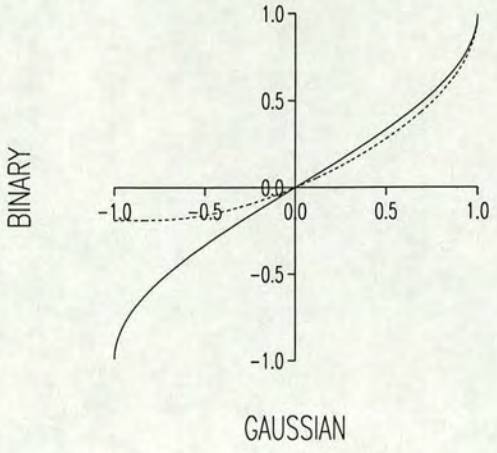

Figure 3.3: *Relationship between expected Gaussian and binary autocorrelation, with threshold level at (—) 0, and (- - -) 1 standard deviation.*

$$= \frac{(\Phi_{T,T}(\rho_l^{(G)}) - 2\Phi_T + 1) - (1 - \Phi_T)^2}{(1 - \Phi_T) - (1 - \Phi_T)^2}$$
$$= \frac{\Phi_{T,T}(\rho_l^{(G)}) - \Phi_T^2}{\Phi_T - \Phi_T^2}. \tag{3.4}$$

where γ_l is the autocovariance at lag l, Φ_T is the cumulative distribution function of the standard normal distribution and $\Phi_{T,T}(\rho_l^{(G)})$ is the cumulative distribution function of the bivariate standard normal distribution with correlation coefficient $\rho_l^{(G)}$, i.e.

$$\Phi_{T,T}(\rho_l^{(G)}) = \int_{y_t=-\infty}^{\hat{T}} \int_{y_{t+l}=-\infty}^{\hat{T}} f(y_t, y_{t+l}; \rho_l^{(G)}) dy_t dy_{t+l}.$$

Equation (3.4) results from (3.3) using identities relating to the bivariate normal probability function, as given in Abramowitz and Stegun (1965, page 936, section 26.3), or Johnson and Kotz (1972, Chapter 36, page 94). This equation can also be used to transform from the sample autocorrelation of the binary time series, denoted $\hat{\rho}_l^{(B)}$ at lag l, to the sample autocorrelation of the latent Gaussian process, $\hat{\rho}_l^{(G)}$, the tetrachoric correlations. Alternatively, by consideration of sample values in (3.3) we get

$$\hat{\rho}_l^{(B)} = \frac{\frac{d}{n} - \left(\frac{b+d}{n}\right)^2}{\frac{b+d}{n} - \left(\frac{b+d}{n}\right)^2}.$$

Figure 3.3 shows the relationship between the expected Gaussian and binary autocorrelation for two levels of thresholding. The threshold levels for the individual cows are given later (in Table 3.4).

3.3.3 Allowing for time trend

Everything above assumes a stationary underlying series and a constant probability of y_t being above the threshold. I now consider the case for which the probability of being above the threshold is no longer constant. This might be due to an overall time trend, or to a diurnal or seasonal effect. The Gaussian series can now be assumed to be non-stationary or, equivalently, the series can be assumed to be stationary but with a threshold that varies with time. Parametrically, the probability of being above the threshold would be modelled via the logit transformation using a linear or sinusoidal function.

If no obvious parametric form for the trend is adequate, as we find for the diurnal cycle in the feeding data in Section 3.7, an alternative approach is to estimate the probability of feeding at a particular time of day by averaging observations at nearby times for all days. Cross-validation can be used to select an optimal window width by omitting each day's data in turn and then predicting it. Then, instead of considering the autocorrelation of the Gaussian process as in Section 3.3.2, the estimate at lag l can be considered as the result of the maximisation of a quasi-log-likelihood of the form

$$\sum_t \log \left[\int_{I_t} \int_{I_{t+l}} f(y_t, y_{t+l}; \rho_l^{(G)}) \, dy_t \, dy_{t+l} \right]. \tag{3.5}$$

Here, the integration interval, I_t, is $(-\infty, \hat{T}_t)$ if $x_t = 0$ and $(\hat{T}_t, \infty)$ if $x_t = 1$. Instead of the previous constant threshold $\hat{T}$, we now have $\hat{T}_t$, chosen so that the probability of not feeding, Φ_{T_t}, matches the diurnal trend. So for a given time of day t, we use some method to estimate the probability that $x_t = 0$, and then calculate $\hat{T}_t = \Phi^{-1}(P(x_t = 0))$.

In the absence of trend, (3.5) can be reduced to

$$\begin{aligned} & a \log \left[\int_{-\infty}^{\hat{T}} \int_{-\infty}^{\hat{T}} f(y_t, y_{t+l}; \rho_l^{(G)}) \, dy_t \, dy_{t+l} \right] \\ + \; & 2b \log \left[\int_{-\infty}^{\hat{T}} \int_{\hat{T}}^{\infty} f(y_t, y_{t+l}; \rho_l^{(G)}) \, dy_t \, dy_{t+l} \right] \\ + \; & d \log \left[\int_{\hat{T}}^{\infty} \int_{\hat{T}}^{\infty} f(y_t, y_{t+l}; \rho_l^{(G)}) \, dy_t \, dy_{t+l} \right], \end{aligned}$$

and $\rho_l^{(G)}$ simplifies to the tetrachoric correlation, as given by the solution of (3.2), or, equivalently, by the relationship (3.4).

3.4 Fast methods for parameter estimation

In this section, I describe several ad-hoc methods of parameter estimation, all of which involve matching the sample values of the autocorrelation to their expected values in some way. For a series of length n, assuming circularity, we have a full set of $n/2$ sample correlations as in (3.1). As ARMA processes only have short-term memory, most of the useful information about the process is contained in the first few lags, and the higher-lag sample correlations are mostly noise. Therefore in any matching of sample correlations to their expected values, we might expect to still get good results by consideration of only the first few lags, i.e. replace n by some $n' < n$. Therefore in all the methods described, it is of interest to compare results for $n' = 2, 4, \ldots, n$.

3.4.1 Ordinary least squares (OLS)

Perhaps the most obvious way to estimate ARMA parameters from the sample autocorrelation coefficients is via least squares, for example as done by Glasbey et al. (1998). We simply want to find the values of the parameters that minimise

$$\sum_{l=1}^{n'/2} (\hat{\rho}_l - \rho_l)^2$$

using either binary or Gaussian autocorrelations and some choice of n'.

3.4.2 Weighted least squares (WLS)

Weighted and generalised least squares are natural methods to consider as improvements over ordinary least squares. However analytical forms cannot be derived for the variances and covariances of either the tetrachoric or binary autocorrelation coefficients. But with the tetrachoric autocorrelation estimating the Gaussian autocorrelation, this suggests that we might be able to use the variances of the Gaussian autocorrelation coefficients as the weights for the tetrachoric autocorrelation. Therefore we minimise

$$\sum_{l=1}^{n'/2} \frac{1}{w_l} (\hat{\rho}_l - \rho_l)^2, \tag{3.6}$$

where $\hat{\rho}_l$ and ρ_l are the sample and expected autocorrelations at lag l, respectively. The weights are given by (see for example Kendall et al., 1983, page 548):

$$w_l = \mathrm{Var}(\rho_l) = \frac{1}{n} \sum_{i=-\infty}^{\infty} \left(\rho_i^2 + \rho_{i-l}\rho_{i+l} - 4\rho_l\rho_i\rho_{i+l} + 2\rho_i^2\rho_l^2 \right)$$

for $l = 1, \ldots, n/2$.

For an AR(1) process we have $\rho_l = \phi^{|l|}$ (and $\rho_{-l} = \rho_l$). Therefore, for $l \geq 1$ we have

$$w_l = \frac{1}{n(1-\phi^2)} \left(1 + \phi^2 - \phi^{2l}[1 + 2l - \phi^2(2l-1)]\right). \tag{3.7}$$

For an ARMA(1,1) process this generalises to

$$\begin{aligned} w_l &= \frac{1}{n(1-\phi^2)} \left(1 + (2\rho_1^2 - \phi^2) - \rho_1\phi^{2l-3}[4(l-1)\rho_1^2 - (2l-7)\rho_1\phi - 2\phi^2]\right. \\ &- \left. \rho_1\phi^{2l-2}[4\rho_1^3 + 4(l-3)\rho_1^2\phi - (2l-9)\rho_1\phi^2 - 2\phi^3]\right), \end{aligned} \tag{3.8}$$

where

$$\rho_1 = \frac{(1-\phi\theta)(\phi-\theta)}{(1+\theta^2-2\phi\theta)}.$$

For an MA(1) process we have no use for weighted (or generalised) least squares, because the expected value of all autocorrelation coefficients at lags greater than 1 is zero.

There is also a determinant term associated with (3.6) which is a function of the parameters, and so including this would give us a method which maximises the log-likelihood, i.e. minimise

$$\sum_{l=1}^{n'/2} \left(\log(w_l) + \frac{1}{w_l}(\hat{\rho}_l - \rho_l)^2\right)$$

which up to a constant is equal to $-2 \times$ log-likelihood. The effect of including this extra term in the minimisation was found to be negligible, and also increased computation time considerably, hence this modification will not be considered further.

The weights as given assume we are dealing with correlations, which is the case here as we can reconstruct the underlying Gaussian variable with unit variance. If we were dealing with a half-censored case, such as rainfall data, for which the variable is observed when above the threshold, we might choose to deal with covariances instead of correlations and then different weight formulae apply. See for example Kendall et al. (1983, page 548).

3.4.3 Generalised least squares (GLS)

Again there are no analytic forms for the covariances of the tetrachoric correlations themselves and so again we consider the corresponding expressions for the

Gaussian autocorrelation for use as the weights. We consider minimising

$$(\hat{\rho} - \rho)^T W^{-1} (\hat{\rho} - \rho),$$

where $\hat{\rho}$ and ρ are vectors of length $n'/2$ containing the sample and expected correlations up to lag $n'/2$. The weight matrix W $(n'/2 \times n'/2)$ is the covariance matrix for the correlations, where elements are given by

$$\begin{aligned} W_{kl} &= \mathrm{Cov}(\rho_k, \rho_l) \\ &= \frac{1}{n} \sum_{i=-\infty}^{\infty} \left(\rho_i \rho_{i+l-k} + \rho_{i-k} \rho_{i+l} - 2\rho_l \rho_i \rho_{i+k} - 2\rho_k \rho_i \rho_{i+l} + 2\rho_i^2 \rho_k \rho_l \right) \end{aligned}$$

for $k, l = 1, \ldots, n/2$. This formula is quoted incorrectly in Kendall et al. (1983, page 573), copied directly from Bartlett (1946), where it was later corrected.

For an AR(1) process we have

$$\begin{aligned} W_{kl} &= \frac{1}{n(1-\phi^2)} \left(\phi^{l-k} [(l-k+1) - \phi^2 (l-k-1)] \right. \\ &+ \left. \phi^{k+l} [(k+l+1) - \phi^2 (k+l-1)] \right) \end{aligned}$$

for $l > k \geq 1$ and $l - k \geq 1$, and the diagonal elements W_{ll} are given by the w_l derived for weighted least squares above (3.7).

For an ARMA(1,1) process this generalises to

$$\begin{aligned} W_{kl} &= \frac{1}{n(1-\phi^2)} \left(\rho_1 \phi^{l-k-2} \left[(l-k-1)\rho_1 + 2\phi - \phi^2 [(l-k-3)\rho_1 + 2\phi] \right] \right. \\ &+ \left. \rho_1 \phi^{k+l-2} \left[(k+l-1)\rho_1 + 2\phi - \phi^2 [(k+l-3)\rho_1 + 2\phi] \right] \right) \end{aligned}$$

for $l > k \geq 1$ and $l - k \geq 1$, and again the diagonal elements W_{ll} are given by the w_l for weighted least squares (3.8).

Inclusion of the determinant term results in the minimisation of

$$\log |W| + (\hat{\rho} - \rho)^T W^{-1} (\hat{\rho} - \rho),$$

but again this gives no obvious benefit in efficiency and increases computation time substantially, and so we will not consider it further.

3.4.4 Pairwise likelihood

For applications in spatial statistics, computation of full likelihoods is often not possible. For variogram estimation, forms of weighted and generalised least

squares can be considered, for example see Cressie (1985). An alternative approach, that of pairwise likelihood, is considered for covariance estimation by Hjort and Omre (1994, pages 305–307), who call this method of maximising a product of bivariate likelihoods *quasi-likelihood*, since it depends only on second-order properties of series. Heagerty and Lele (1998) use this pairwise likelihood approach in the context of spatial probit regression, and Nott and Rydén (1999) go on to consider a weighted form.

The form of the pairwise likelihood we consider here is simply the product of all bivariate probabilities for pairs of observations, (y_t, y_{t+l}), with $0 < l \leq n'/2$,

$$Q = \prod \frac{1}{\sqrt{2\pi(1-\rho_l^{(G)2})}} \exp\left[-\frac{1}{2}\frac{y_t^2 + y_{t+l}^2 - 2\rho_l^{(G)} y_t y_{t+l}}{1-\rho_l^{(G)2}}\right].$$

This leads to the minimisation of

$$\sum_{l=1}^{n'/2} \log(1-\rho_l^{(G)2}) + 2\sum_{l=1}^{n'/2}\left(\frac{1-\hat{\rho}_l^{(G)}\rho_l^{(G)}}{1-\rho_l^{(G)2}}\right).$$

In this form it can be seen that this too is just an alternative method of matching the sample autocorrelation coefficients with their expected values.

3.4.5 Spectral likelihood

Sample autocorrelation coefficients at different lags are highly correlated, so the methods described so far are not necessarily efficient estimation procedures. An alternative is to transform to independent statistics, for which the natural choice is by the Fourier transform. Glasbey et al. (1998) considered this idea for full series, here we consider the same technique applied to censored series.

Whittle (1953) derived the spectral approximation for the log-likelihood, $\mathcal{L}$, of an R-dimensional stationary multivariate Gaussian process of length n. This is considered in detail in Chapter 4, both in the univariate case and in the multivariate case, and I prove that the likelihood written in this spectral form may be approximated by a restricted form by taking $n' < n$. The rationale behind taking $n' < n$ is the same as discussed previously. Assuming the circulant model, a series of length n has $n/2$ sample autocorrelation coefficients. These lead to $(n/2)+1$ independent periodogram coefficients. For a short-term memory process, alternative periodogram coefficients can be formed from just the lower lags, with little loss of information. As before, this may be conceptually intuitive, but it is not mathematically intuitive here, since this 'restricted' periodogram corresponds to a different set of frequencies than the full periodogram.

The main details in the univariate case are outlined here, to avoid reference to Chapter 4. Note that for the application to feeding data, as the latent process is never observed, it can be assumed, without loss of generality, that the latent process has unit variance, and hence the distinction between (inverse) autocovariance and autocorrelation is unimportant. In Chapter 4 the results are presented in the more general form using covariances; in this chapter, for simplicity, correlations are used.

The full log-likelihood for a stationary, mean-corrected Gaussian time series is given by

$$\mathcal{L} = -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log S_k - \frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \frac{\hat{S}_k}{S_k}. \tag{3.9}$$

Here S_k and $\hat{S}_k$ are, respectively, the spectral and periodogram coefficients at frequency $2\pi k/n$, so

$$\hat{S}_k = \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \hat{\rho}_l^{(G)} e^{\frac{-2\pi i k l}{n}} \qquad \text{for } k = -\frac{n}{2}, \ldots, \frac{n}{2} - 1, \tag{3.10}$$

with S similarly defined in terms of $\rho^{(G)}$. Note that as the Gaussian process is latent, (3.9) can only be considered as a quasi-likelihood, so we can also consider the same functional expression but with $\rho^{(G)}$ replaced by $\rho^{(B)}$.

Glasbey et al. (1998) replace n by n' in (3.9), to obtain a *restricted* log-likelihood, $\mathcal{L}'$, given by

$$\mathcal{L}' = -\frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log|S'_k| - \frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \frac{\hat{S}'_k}{S'_k},$$

where S' and $\hat{S}'$ are obtained as the discrete Fourier transforms of cross-correlations up to lag $n'/2$ only, by replacing n by n' in (3.10). They show empirically, using uncensored AR(1) and ARMA(1,1) models as examples, that for sufficiently large n', maximisation of $\mathcal{L}'$ produces the same parameter estimates as maximisation of the full $\mathcal{L}$. In Chapter 4, we prove that for short-memory processes such as ARMA models, $\mathcal{L}' \approx \mathcal{L}$ for sufficiently large n'. In practice we can take $n' \ll n$, which leads to considerable computational saving. The main conditions needed for the approximation are that S_k is a continuous function of k and that $\rho_l^{(G)}$ and α_l are negligible for $|l| > n'/2$, where α_l is the *inverse autocorrelation coefficient* at lag l, defined as

$$\alpha_l = \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \frac{1}{S_k} e^{\frac{2\pi i k l}{n}} \qquad \text{for } l = -\frac{n}{2}, \ldots, \frac{n}{2} - 1,$$

as considered by Cleveland (1972) and Chatfield (1979).

	AR(1) $\phi = 0.6$			MA(1) $\theta = -0.3$			ARMA(1,1) $(\phi, \theta) = (0.9, 0.6)$		
n'	$\rho_{n'/2}$	$\alpha_{n'/2}$	$\mathcal{L}'$	$\rho_{n'/2}$	$\alpha_{n'/2}$	$\mathcal{L}'$	$\rho_{n'/2}$	$\alpha_{n'/2}$	$\mathcal{L}'$
2	0.60	−0.44	391.119	0.28	−0.30	481.004			
4	0.36	0.00	282.468	0.00	0.090	456.386	0.44	−0.11	373.575
10	0.078	0.00	281.388	0.00	−0.0024	458.353	0.32	−0.025	329.603
20	0.0060	0.00	282.237	0.00	$< 10^{-5}$	458.356	0.19	−0.0019	317.364
50	$< 10^{-5}$	0.00	282.239	0.00	$< 10^{-13}$	458.356	0.039	$< 10^{-6}$	315.662
100	$< 10^{-11}$	0.00	282.239	0.00	$< 10^{-15}$	458.356	0.0028	$< 10^{-11}$	315.770
500	$< 10^{-15}$	0.00	282.239	0.00	$< 10^{-15}$	458.356	$< 10^{-11}$	$< 10^{-15}$	315.775
1000	$< 10^{-15}$	0.00	282.239	0.00	$< 10^{-15}$	458.356	$< 10^{-15}$	$< 10^{-15}$	315.775

Table 3.1: *Values of $\mathcal{L}'$, for a range of n', at true parameter values, averaged over 100 simulated series of length 1000. The expected values of autocorrelation and inverse-autocorrelation coefficients are shown for lag $n'/2$, i.e. $\rho_{n'/2}$ and $\alpha_{n'/2}$ respectively. Where upper bounds are shown, this is for magnitude only, the sign being omitted.*

These conditions hold for ARMA processes, typically for small values of n', because autocorrelations and inverse autocorrelations decay exponentially (Chatfield, 1979; Box et al., 1994, page 79). The rate of convergence, illustrated in Table 3.1, depends on the rate of decay of ρ_l and α_l. Typically, the likelihoods in the table are approximated to within ± 0.0005 of their true value if n' is such that ρ_l and α_l are of the order 10^{-4} for $l > n'/2$.

3.5 MCMC methods

All the methods described in Section 3.4 are computationally fast, especially when $n' \ll n$. Their relative efficiency can be compared by use of the root mean square error (RMSE), but it would also be useful to have an idea of how these compare to the maximum efficiency attainable in theory. A Markov chain Monte-Carlo (MCMC) approach is therefore employed to see what efficiency is attainable with such a computer-intensive method. The main details are given in Section 3.5.1, before giving detailed methodological details for each of the process types AR(1), MA(1) and ARMA(1,1) in Sections 3.5.2–3.5.4.

3.5.1 General methodology

Gibbs sampling is used to simulate realisations of the latent Gaussian series consistent with the thresholding dictated by the binary series. For each realisation of the complete series, estimates of the parameters are obtained by sampling from the likelihood via a Metropolis-Hastings step. Finally, we average these estimates over a large number of iterations.

The general methodology is essentially the same for the three types of process

considered — AR(1), MA(1) and ARMA(1,1). Differences occur only in the complexity of the Gibbs sampling, AR(1) processes being the simplest to deal with, and extra steps needing to be added in going to MA(1) and again in going to ARMA(1,1). For AR(1), the conditional distribution of $Y_t|(Y_{-t} = y_{-t})$ (where Y_{-t} is used to indicate the whole series Y but omitting Y_t) is simply the conditional distribution $Y_t|(Y_{t-1} = y_{t-1}, Y_{t+1} = y_{t+1})$, as the inverse covariance matrix is tri-diagonal. The distributional form is easily derived and sampling is straightforward. For MA(1), the conditional distribution involves the whole series and so calculations would involve the inversion of an $n \times n$ matrix. This can be overcome using methods described by Phadke and Kedem (1978) and Ansley (1979), using the Cholesky decomposition to transform the original series into a set of independent observations for which the likelihood is easily evaluated. For ARMA(1,1), the use of a further transformation enables the same methodology to be used. Luceño (1993) also discusses such methodologies and considers efficient ways of arranging computations. For example, for an AR(1) process, the variance, the correlation at lag 1 and the sum of squared first and last observations form a set of sufficient statistics for ϕ and therefore once these have been calculated the likelihood can be computed for different estimates of ϕ by just two multiplications and two additions. This can be utilised when comparing the likelihood with the current value of ϕ and the new candidate for ϕ, but obviously for the next realisation of the chain, the sufficient statistics must be recalculated.

3.5.1.1 Convergence of the Markov chain

Many techniques have been developed to assess whether a Markov chain can be considered to have converged; a comprehensive review is given in Brooks and Roberts (1998). The package CODA (Best et al., 1995, 1997) was used to assess the chains using some of these methods. Simple statistics and plots of the chains were examined, and diagnostics due to Geweke and to Heidelberger and Welch were used to check convergence more formally (see Appendix B for details of these tests). From consideration of these I decided that a burn-in of 500 iterations followed by a further 10000 realisations was sufficient to assume convergence of the chains. For a given series, single estimates of the parameters were calculated as the mean of the 10000 realisations.

Some examples of CODA output are given in Appendix B, which shows the above-mentioned plots, statistics and tests from four simulated series of the following types:

- AR(1) process with $\phi = 0.6$, thresholding at 1 standard deviation and series length 1000,
- MA(1) process with $\theta = -0.3$, thresholding at 1 standard deviation and series length 1000.

3.5.2 Methodology for an AR(1) process

Here we present full details of the methodology used for AR(1) processes.

- From the binary series $x = (x_0, x_1, \ldots, x_{n-1})$, an arbitrary method is used to create an initial Gaussian series $y^{(0)} = (y_0^{(0)}, y_1^{(0)}, \ldots, y_{n-1}^{(0)})$ such that if $x_t = 0$ then $y_t^{(0)} < \hat{T}$ and if $x_t = 1$ then $y_t^{(0)} > \hat{T}$, where $\hat{T}$ is the threshold of the latent Gaussian series that corresponds to the binary series.

  The method employed to obtain the initial Gaussian series is

$$y_t^{(0)} \sim \begin{cases} N(\hat{T} - 0.1,\ 0.025^2) & \text{if } x = 0 \\ N(\hat{T} + 0.1,\ 0.025^2) & \text{if } x = 1. \end{cases}$$

  The value of $y_t^{(0)}$ is re-simulated if it falls the wrong side of the threshold (unlikely with the values chosen).

- From this series, $\phi^{(0)}$ is set equal to the maximum likelihood estimate.

  The log-likelihood for a general multivariate normal series y is given by $\mathcal{L}$, where

$$-2\mathcal{L} = \log|V| + y^T V^{-1} y + \text{constant}, \tag{3.11}$$

  where V is the variance matrix with elements $V_{kl} = \gamma_{|k-l|}$.

  For an AR(1) process, V^{-1} is a band matrix with elements $(0 \leq k, l \leq n-1)$

$$\left(V^{-1}\right)_{kl} = \begin{cases} 1 & \text{if } k = l = 0 \text{ or } k = l = n-1 \\ 1 + \phi^2 & \text{if } k = l \text{ and } 1 \leq k \leq n-2 \\ -\phi & \text{if } |k-l| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

  Using this, and omitting the constant, we can write the log-likelihood as $\mathcal{L}$, where

$$-2\mathcal{L} = -\log(1-\phi^2) + y_0^2 + (1+\phi^2)\sum_{t=1}^{n-2} y_t^2 + y_{n-1}^2 - 2\phi\sum_{t=0}^{n-2} y_t y_{t+1}. \tag{3.12}$$

  $\mathcal{L}$ is then maximised to get an estimate $\hat{\phi}$. If $\hat{\phi} = \pm 1$ we set $\phi^{(0)} = \pm 0.9999$, otherwise we set $\phi^{(0)} = \hat{\phi}$. It should be noted that estimating ϕ by the

lag 1 autocorrelation is *not* the maximum likelihood estimate, which would instead correspond to ignoring the determinant term in the likelihood of (3.12).

- A series $y^{(i)}$ is created from $y^{(i-1)}$ for $i = 1, 2, \ldots$ by replacing each $y_t^{(i-1)}$ in turn with $y_t^{(i)}$. In general this is done by replacing with a value simulated from the full conditional distribution, i.e.

$$Y_t|Y_{-t} = Y_t|(Y_0 = y_0, Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}, Y_{t+1} = y_{t+1}, \ldots, Y_{n-1} = y_{n-1}).$$

In the case of an AR(1) process this simplifies to:

$$Y_t|(Y_{-t} = y_{-t}) = Y_t|(Y_{t-1} = y_{t-1}, Y_{t+1} = y_{t+1}).$$

Using Tong (1990, section 3.3), this conditional distribution is normal, and the mean and variance are derived as

$$\begin{aligned} \mathrm{E}\left[Y_t|(Y_{-t} = y_{-t})\right] &= \frac{\phi}{1+\phi^2}(y_{t-1} + y_{t+1}) \\ \mathrm{Var}\left[Y_t|(Y_{-t} = y_{-t})\right] &= \frac{1}{1+\phi^2} \end{aligned}$$

for $t = 1, \ldots, n-2$.

For the first and last observations in the series, the conditional distributions are given by

$$\begin{aligned} Y_0|(Y_1 = y_1) &\sim N(\phi y_1, 1) \\ Y_{n-1}|(Y_{n-2} = y_{n-2}) &\sim N(\phi y_{n-2}, 1). \end{aligned}$$

Updates are simulated from the parts of the above distributions which correspond to the correct side of the threshold as dictated by x_t.

Hence in detail, the sub-steps for this step are as follows.

- simulate ${y_0}^{(i)}$ from the conditional distribution $Y_0|\left(Y_1 = {y_1}^{(i-1)}\right)$,
- simulate ${y_1}^{(i)}$ from $Y_1|\left(Y_0 = {y_0}^{(i)}, Y_2 = {y_2}^{(i-1)}\right)$
- simulate ${y_2}^{(i)}$ from $Y_2|\left(Y_1 = {y_1}^{(i)}, Y_3 = {y_3}^{(i-1)}\right)$

  $\vdots$

- simulate ${y_{n-2}}^{(i)}$ from $Y_{n-2}|\left(Y_{n-3} = {y_{n-3}}^{(i)}, Y_{n-1} = {y_{n-1}}^{(i-1)}\right)$
- simulate ${y_{n-1}}^{(i)}$ from $Y_{n-1}|\left(Y_{n-2} = {y_{n-2}}^{(i)}\right)\Big)$.

- From the resulting series y^i, a new estimate $\phi^{(i)}$ is obtained using a Metropolis-Hastings step.
  - The maximum likelihood estimate $\hat{\phi}$ is calculated by maximising $\mathcal{L}$, as given by (3.12).
  - Using asymptotic theory for maximum likelihood estimation, (see for example Stuart et al., 1999, page 60), we estimate the standard error of this estimator to be

$$\begin{aligned} \text{se}(\hat{\phi}) &\approx 1 \Big/ \left[-\text{E}\left(\frac{\partial^2 \mathcal{L}}{\partial \phi^2} \right) \right]_{\phi=\hat{\phi}} \\ &= \frac{(1-\hat{\phi}^2)}{\sqrt{n(1-\hat{\phi}^2)+3\hat{\phi}^2-1}} \\ &\approx \sqrt{\frac{1-\hat{\phi}^2}{n}} \quad \text{for large } n, \end{aligned}$$

    where $\mathcal{L}$ is the log-likelihood as given in (3.12).
  - A candidate ϕ' for $\phi^{(i)}$ is simulated from proposal distribution $f(\phi) = N(\hat{\phi}, \text{se}(\hat{\phi})^2)$.
  - Set $\phi^{(i)} = \phi'$ with probability given by

$$\min \left\{ 1, \frac{l(\phi')}{l(\phi^{(i-1)})} \frac{f(\phi^{(i-1)})}{f(\phi')} \right\}$$

    where $f(\cdot)$ is our asymptotic approximation to the likelihood that we have simulated from, and $l(\cdot)$ is the exact likelihood, the log of which is given by (3.12). Otherwise set $\phi^{(i)} = \phi^{(i-1)}$.

These steps are repeated a large number of times to create the Markov chain. For a given simulated binary series, a single estimate of ϕ is calculated as the mean of the last 10000 values. The whole procedure is performed on 100 simulated series and a RMSE calculated from these 100 means. It can be noted that the posterior mean is the optimal estimator for MSE loss.

3.5.3 Methodology for an MA(1) process

For MA(1) processes, the full conditional distribution needed for the Gibbs sampling involves the whole series. Therefore in order to avoid having to perform the inversion of an $n \times n$ matrix, the Cholesky decomposition, see for example Tong (1990, page 184), is used to transform the original series into a set of independent observations for which the likelihood is easily evaluated.

- From the binary series, an initial Gaussian series is created in the same way as was done in the AR(1) case.

- $\theta^{(0)}$ is set equal to the maximum likelihood estimate for this series. However, evaluating the likelihood for an MA(1) process involves the non-trivial inversion of an $n \times n$ matrix, where n is the series length, so to avoid this we use a method described by Phadke and Kedem (1978) and later by Ansley (1979) which utilises the band structure of the covariance matrix V for an MA(1) process, using the Cholesky decomposition of this to transform our original series y into a series of independent observations w, for which the likelihood is easily evaluated. In detail, the following steps are carried out.

  - The log-likelihood is again given by (3.11). For MA(1) processes, V is a band matrix with elements

$$V_{kl} = \begin{cases} 1+\theta^2 & \text{if } k = l \\ -\theta & \text{if } |k-l| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

    The matrix V has the Cholesky decomposition $V = CC^T$ where C is lower triangular with non-zero elements only on the leading diagonal and the first sub-diagonal. It is then easy to solve the equation $w = C^{-1}y$. The sequence given by $w = (w_0, w_1, \ldots, w_{n-1})$ is a set of independent normal variables with zero mean and unit variance. The Jacobian of the transformation $y \mapsto w$ is 1 and hence, omitting the constant, the log-likelihood (3.11) can be written

$$\begin{aligned} -2\mathcal{L} &= 2\log|C| + w^T w \\ &= 2\sum_{t=0}^{n-1} \log C_{tt} + \sum_{t=0}^{n-1} w_t^2. \end{aligned} \tag{3.13}$$

- The next step is to create series $y^{(i)}, i = 1, 2, 3, \ldots$, replacing each $y_t^{(i-1)}$ in turn. As for AR(1) this is done by Gibbs sampling, but here the inverse variance matrix is not a band matrix, so the Cholesky decomposition will again be used to obtain the conditional distribution.

  - Using Tong (1990, section 3.3) as before, we obtain the conditional distribution of $Y_t|(Y_{-t} = y_{-t})$ by partitioning Y as

$$Y = \begin{pmatrix} Y_t \\ Y_{-t} \end{pmatrix},$$

    where $Y \sim N_n(0, V)$, with

$$\mu = \begin{pmatrix} \mu_t \\ \mu_{-t} \end{pmatrix} = \mathbf{0}, \quad V = \begin{pmatrix} V_{t,t} & V_{t,-t} \\ V_{-t,t} & V_{-t,-t} \end{pmatrix}.$$

Then the conditional distribution is normal with mean and variance respectively given by

$$\begin{aligned} \mu_C &= V_{t,-t}\, V_{-t,-t}^{-1}\, y_{-t} \\ V_C &= V_{t,t} - V_{t,-t}\, V_{-t,-t}^{-1}\, V_{-t,t}. \end{aligned}$$

Then using the Cholesky decomposition $V_{-t,-t} = C_{-t,-t}\, C_{-t,-t}^T$ and solving

$$\begin{aligned} S_{-t,t} &= C_{-t,-t}^{-1}\, V_{-t,t} \\ \text{and } w_{-t} &= C_{-t,-t}^{-1}\, y_{-t}, \end{aligned}$$

we have

$$\begin{aligned} \mu_C &= S_{-t,t}^T\, w_{-t} \\ \text{and } V_C &= V_{t,t} - S_{-t,t}^T\, S_{-t,t}. \end{aligned}$$

  - $y_t^{(i)}$ can now be simulated from the part of $N \sim (\mu_C, V_C)$ which corresponds to the correct side of the threshold as determined by x_t.

- Once every element in the series has been updated, we need to simulate a value for $\theta^{(i)}$. For AR(1) we simulated from the asymptotic distribution for the likelihood and used a Metropolis-Hastings algorithm. Here it would involve a lot of computation to obtain the standard error for the MLE, so instead we replace the standard error with a fixed, arbitrary value, set to achieve an acceptance probability of somewhere in the range 15–50% (Gelman et al., 1996). So the steps are as follows.
  - Calculate the MLE $\hat{\theta}$ given the current series.
  - Simulate a candidate θ' for $\theta^{(i)}$ from proposal distribution $f(\theta) = N(\hat{\theta}, s^2)$, with some value of s.
  - This new candidate for $\theta^{(i)}$ is accepted with probability

$$\min\left\{1, \frac{l(\theta')}{l(\theta^{(i-1)})} \frac{f(\theta^{(i-1)})}{f(\theta')}\right\},$$

    where $l(\cdot)$ is the exact likelihood and $f(\cdot)$ is as just defined. Otherwise set $\theta^{(i)} = \theta^{(i-1)}$.

Note that if we set s to a small value, the chain moves around a lot, but has the danger that if it moves to an unlikely value then it can have difficulty getting out of it again, due to $f(\theta^{(i-1)})$ being very small. Conversely, if s is set too large,

many candidates are generated far from the true value and are not accepted, and so the chain will not often move. A value somewhere between these two extremes needs to be selected, and, by trial and error, $s = 0.3$ was decided to be a suitable value here and has been used throughout.

Again chains of length $500 + 10000$ iterations were used and again CODA was used to check adequate convergence. For each simulated series, the final parameter estimate was calculated as the mean of the final 10000 values.

3.5.4 Methodology for an ARMA(1,1) process

We can use essentially the same methodology here as for the MA(1) processes, but need to employ a transformation to allow the ARMA(1,1) process to be treated like an MA(1) process.

- The binary series is taken and an initial Gaussian series created in the same way as in the cases above.
- To evaluate the likelihood the method described by Ansley (1979) is used. Consider the transformation $y \mapsto w$ given by

$$w_t = \begin{cases} y_t & \text{for } t = 0 \\ y_t - \phi y_{t-1} & \text{for } t = 1, \ldots, n-1. \end{cases} \tag{3.14}$$

  Then w_t is a modified MA(1) process, the covariance matrix V of which is given by

$$V_{kl} = \begin{cases} \dfrac{1 + \theta^2 - 2\phi\theta}{1 - \phi^2} & \text{if } k = l = 0 \\ 1 + \theta^2 & \text{if } k = l, k = 1, \ldots, n-1 \\ \dfrac{(1 - \phi\theta)(\phi - \theta) - \phi(1 + \theta^2 - 2\phi\theta)}{1 - \phi^2} & \text{if } |k - l| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

  The Cholesky decomposition can be used as before and the likelihood given by (3.13) maximised, but now with this covariance matrix. $(\phi^{(0)}, \theta^{(0)})$ are set to their maximum likelihood estimates.

- We then need to create series $y^{(i)}, i = 1, 2, 3, \ldots$, replacing each element $y_t^{(i-1)}$ in turn. It is too computationally expensive to do this directly, so instead we use the transformation (3.14) with $\phi = \phi^{(0)}$.

To replace each y_t we see that a single y_t depends on both w_t and w_{t+1} through the equations

$$\begin{aligned} w_t &= y_t - \phi y_{t-1} \\ w_{t+1} &= y_{t+1} - \phi y_t. \end{aligned} \tag{3.15}$$

So we can work out the bivariate normal distribution for $\{w_t, w_{t+1}\}$ conditional on the rest of the w_t's. Then using (3.15) we need to simulate a value for y_t; this corresponds to a line through the bivariate distribution of $\{w_t, w_{t+1}\}$.

Combining the two equations of (3.15) we get

$$y_t = w_t + \phi y_{t-1} = \frac{1}{\phi}(y_{t+1} - w_{t+1}),$$

which we can write as the equation of a straight line in $\{w_t, w_{t+1}\}$, i.e.

$$w_{t+1} = -\phi w_t + (y_{t+1} - \phi^2 y_{t-1})$$

We need to simulate a point on this line. To do this we define two further transformations. Firstly, $w \mapsto u$ is a translation through $(0, -c)$, i.e.

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} w_t \\ w_{t+1} \end{pmatrix} + \begin{pmatrix} 0 \\ -c \end{pmatrix}$$

and secondly $u \mapsto v$ is a rotation through an angle $\psi = \arctan(-\phi)$, i.e.

$$\begin{aligned} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} &= \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sqrt{1+\phi^2}} & \frac{\phi}{\sqrt{1+\phi^2}} \\ \frac{-\phi}{\sqrt{1+\phi^2}} & \frac{1}{\sqrt{1+\phi^2}} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \end{aligned}$$

Then if we have

$$\{w_t, w_{t+1}\} | w_{-[t,t+1]} \sim N_2(\mu'_w, V_w),$$

it follows that

$$\{u_1, u_2\} \sim N_2(\mu'_w + (0, -c)', V_w)$$

and

$$\{v_1, v_2\} \sim N_2(M[\mu'_w + (0, -c)'], MV_w M'),$$

and then the distribution we require to simulate from is the marginal v_1.

Using the inverse transformations $v \mapsto u$ and $u \mapsto w$, we obtain

$$\begin{pmatrix} w_t \\ w_{t+1} \end{pmatrix} = \begin{pmatrix} v_1 \cos\psi \\ -v_1 \sin\psi + c \end{pmatrix}.$$

We need to ensure that values are simulated which result in the new y_t being on the correct side of the threshold. For y_0, we have

$$x_0 = 1 \Rightarrow y_0 > \hat{T} \Rightarrow w_0 > \hat{T}.$$

Now $w_t = v_1 \cos\psi$ and so we need to simulate a $v_1 > \hat{T}/\cos\psi$.

For $y_t, t = 1, 2, \ldots$

$$x_t = 1 \Rightarrow y_t > \hat{T} \Rightarrow w_t + \phi y_{t-1} > \hat{T}.$$

Therefore we need $w_t > \hat{T} - \phi y_{t-1}$ and, since $w_t = v_1 \cos\psi$, we need to simulate a $v_1 > (\hat{T} - \phi y_{t-1})/\cos\psi$.

- Once we have the new realisation of the series, $y^{(i)}$, we need to simulate values for $(\phi^{(i)}, \theta^{(i)})$. This is done in the same way as for MA(1), but here there are two parameters to consider so there is a choice of either simulating candidates as a pair, accepting or rejecting both together, or treating each in turn separately. It was decided that better convergence of the chain was achieved by treating both together as a pair. For the Metropolis-Hastings step we again use proposal distributions that are normal, centred on the MLE, and again we arbitrarily choose standard deviations for this distribution to achieve reasonable mixing and convergence — trial and error showed values of 0.5 to be suitable here and were used throughout.

Again chains of length $500 + 10000$ were used and convergence checked using CODA.

3.6 Simulation

A simulation study was carried out in order to compare the efficiency of the spectral estimator with the other fast estimators and with the MCMC method, and to investigate optimal values of n'. We first give details of the simulation study and then discuss results for a range of ARMA processes.

3.6.1 Methodology

We simulated 100 replicates of each of the following ARMA processes.

- AR(1) - $\phi = 0.0, 0.3, 0.6, 0.9$,
- MA(1) - $\theta = 0.0, -0.3, -0.6, -0.9$,

- ARMA(1,1) – $(\phi, \theta) = (0.3, 0.0), (0.0, -0.6), (0.6, 0.3), (0.6, -0.3), (0.3, -0.3), (0.9, 0.6)$.

All series had zero mean, unit variance, were of length $n = 100$ or 1000, and were thresholded at either 0 or 1 standard deviation. Parameters were estimated for a range of values of n' — 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 30, 34, 40, 50, 60, 70, 80, 90, 100. Note that the highest value of 100 is the maximum value for series length 100 but in theory we could have looked at values up to 1000 for the longer series; this was done, but results are not presented. Computation times increase considerably for some of the methods as n' gets large, and inspection of results shows that in general estimates have become stable by this point.

All the processes simulated from have positive autocorrelation at short lags, as this is a necessary property for the type of data we want to model. The AR(1) models have exponentially decaying autocorrelation, the MA(1) models have positive autocorrelation at lag 1 and zero autocorrelation thereafter. The parameters chosen for the ARMA(1,1) processes all result in decaying positive autocorrelation, obtained by taking values such that $0 \leq \phi < 1$ and $-1 < \theta < \phi$. Within this region, if $\theta > 0$ then the partial autocorrelation function is also positive and decaying; for $\theta < 0$ the modulus of this function is decaying but the sign is oscillating, (see, for example Box et al., 1994, page 82). Note that taking $\phi = \theta$ results in an independent series for all values between ± 1. See Section 3.2 for definitions and some of the properties of the ARMA processes considered. Table 3.1 gives some additional properties for a process of each type.

All work was carried out using Fortran 90. Parameter estimation for AR(1) and MA(1) processes was performed by a grid search to find estimates correct to 4 decimal places. Depending on the method, sums of squares or $-2\times$ log-likelihood were calculated for parameter values $-0.99, -0.98, \ldots 0.99$. Then I let $\tilde{\phi}$ be the value from this set giving the lowest function value, and carried out a grid search on $\tilde{\phi} - 0.0099, \tilde{\phi} - 0.0098, \ldots, \tilde{\phi} + 0.0099$. Then the value giving the lowest function value in this set was the parameter estimate correct to 4 decimal places. For ARMA(1,1) processes, this grid search approach would be inefficient as we have to minimise over both ϕ and θ. Therefore here the NAG optimisation routine E04JAF was used (Numerical Algorithms Group, 1993). This minimises a function of several variables using a quasi-Newton algorithm, allowing simple bounds on parameters and using function values only. Output from this routine includes an *ifail* parameter which indicates whether the routine is confident it has found the global minimum. Where this indicated a 'soft fail', e.g. "conditions for a minimum have not all been met but a lower point could not be found" or "some

doubt about whether the point found is a minimum ...", we have still accepted these estimates, with the philosophy that this is still the best estimate that the given method can provide for the given series.

100 series were simulated for each model. From these, if any series had a sample lag 1 correlation of -1, this particular series was discarded and a replacement series simulated. Such series contain very weak information about the underlying correlation structure and so use of these in the simulation study might unnecessarily distort results. This wasn't however a big problem — it only occurs when a thresholded series is generated that contains no adjacent 1s at all, more likely for short series with high threshold and low correlation. It never occurred with any of the series of length $n = 1000$. For length $n = 100$, the numbers of replaced series were
AR(1)/MA(1) — $\phi = \theta = 0.0$: 8 series;
AR(1) — $\phi = 0.3$, $\phi = 0.9$: 1 series each;
MA(1) — $\theta = 0.3$, $\theta = 0.6$, $\theta = 0.9$: 1 series each;
ARMA(1,1) — $(\phi, \theta) = (0.0, -0.6)$, $(\phi, \theta) = (0.3, 0.0)$, $(\phi, \theta) = (0.6, 0.3)$: 1 series each; $(\phi, \theta) = (0.9, 0.6)$: 3 series.

Where sample correlations were equal to -1 at lags higher than 1, all methods replaced this with the expected value. In the case of least squares methods this is equivalent to omitting these terms.

Throughout the study, the same simulated series were used for the two levels of thresholding and for estimation using all methods, including MCMC.

3.6.1.1 Criterion for comparison of estimators

The obvious criterion to compare the efficiency of estimators is the mean square error (MSE) which, for an estimator $\tilde{\phi}$ for ϕ, can be expressed as

$$\text{MSE}(\tilde{\phi}) = \frac{1}{100}\sum_{i=1}^{100}(\tilde{\phi}_i - \phi)^2 = \text{Bias}(\tilde{\phi})^2 + \text{Var}(\tilde{\phi}).$$

For the ARMA(1,1) models with two parameters, we consider the combined MSE, given by the sum of the two individual MSEs. Where possible we express efficiencies relative to that of the MCMC estimator, which we consider as being fully efficient, i.e.

$$\text{Relative efficiency of fast estimator} = \frac{\text{MSE for MCMC estimator}}{\text{MSE for fast estimator}},$$

where both MSEs may be averages over parameter values and threshold levels.

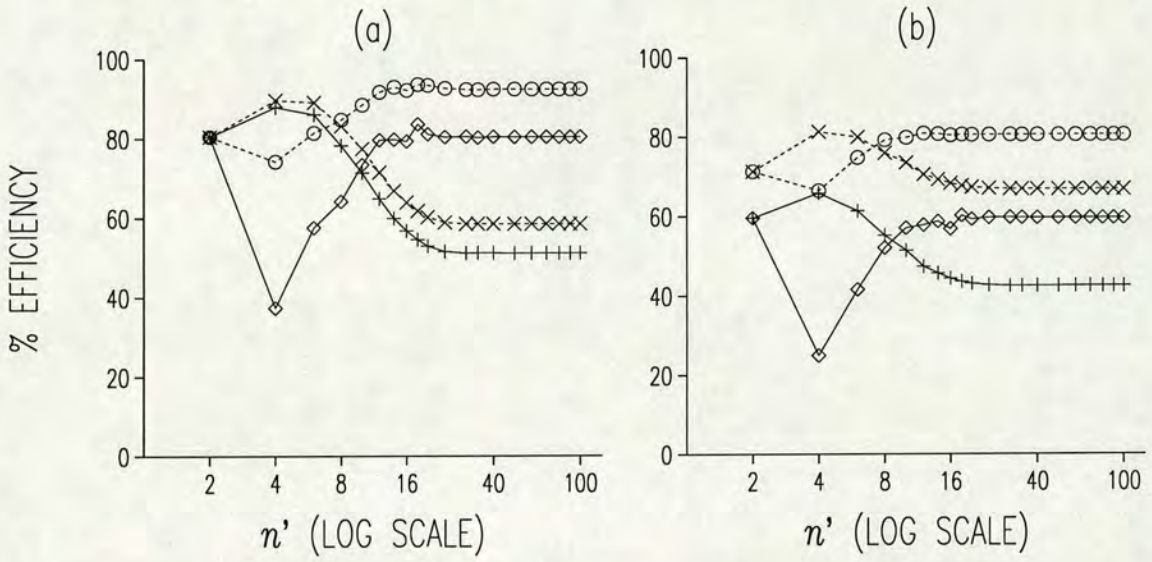

Figure 3.4: *Percentage efficiencies relative to MCMC for estimation in an AR(1) process of length 1000 with $\phi = 0.6$, for a range of values of n'; (a) threshold = 0, (b) threshold = 1 sd; (×- - -×) OLS using binary autocorrelation, (+—+) OLS using Gaussian autocorrelation, (○- - -○) spectral method using binary autocorrelation, (◇—◇) spectral method using Gaussian autocorrelation.*

3.6.2 Simulation results

Appendix C presents detailed simulation results for the whole range of processes considered, for all the fast estimation methods considered in Section 3.4 and, where results are available, for the MCMC method of Section 3.5. Here we discuss a few cases in detail before presenting summaries of the results and discussing the main features.

Figures 3.4 and 3.5 show typical relationships between efficiency and n' for parameter estimation by ordinary least squares and the spectral method using both binary and Gaussian autocorrelation. Figure 3.4 presents results for an example AR(1) process, showing results in terms of % efficiency compared with the MCMC method. Figure 3.5 presents results for an example ARMA(1,1) process and, as MCMC results were not produced for this case due to the large amount of computation required, results are presented as root mean square errors. These are joint RMSEs as described in the last section, calculated as the square root of the sum of the MSEs for ϕ and θ individually. The other methods — weighted and generalised least squares and the pairwise likelihood method — are not shown on the graphs in order to preserve clarity.

Figure 3.4 illustrates well one of the main advantages of the spectral method

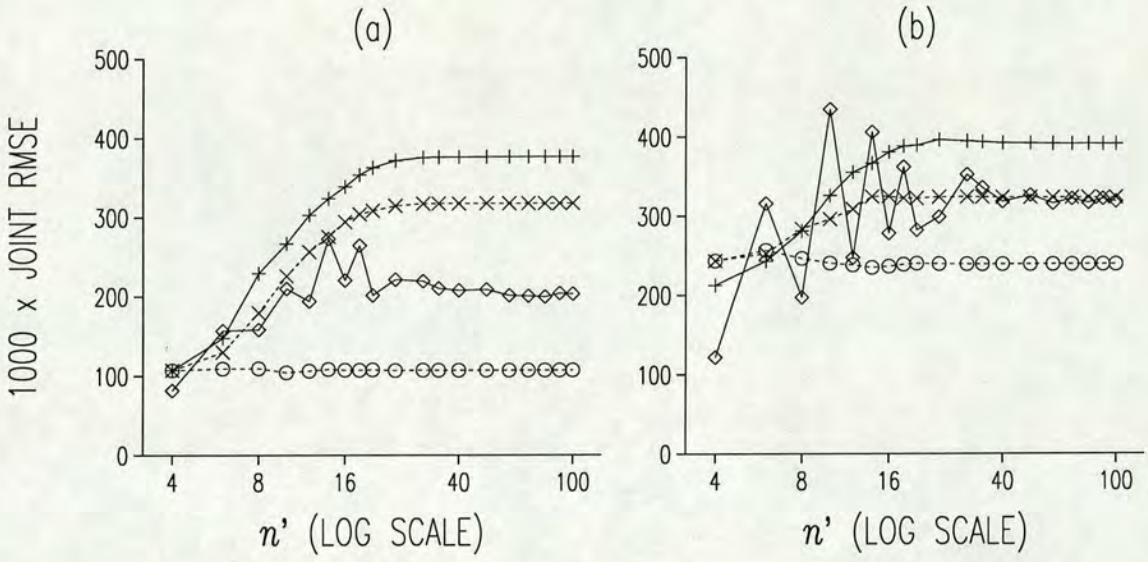

Figure 3.5: *Joint root mean square errors for estimation in an ARMA(1,1) process of length 1000 with* $(\phi, \theta) = (0.6, -0.3)$*, for a range of values of* n'*; (a) threshold = 0, (b) threshold = 1 sd; (×- - -×) OLS using binary autocorrelation, (+— +) OLS using Gaussian autocorrelation, (○- - -○) spectral method using binary autocorrelation, (◇—◇) spectral method using Gaussian autocorrelation.*

over the least squares methods. The optimal value of n' for least squares is quite crucial, here 4, but not known *a priori*, whereas for the spectral method the exact choice of n' is not important — it simply has to be sufficiently large, here larger than 20 is sufficient. This is the general pattern, and so even if the spectral method is not much more efficient than other methods, the real advantage is the lack of needing to know an optimal value for n' in advance. The value needed can be gauged from the rate of decay of the autocorrelation and inverse autocorrelation coefficients (see Section 3.4.5).

Figure 3.5 shows the spectral method using the binary autocorrelation to be stable even for very small values of n'. In contrast, use of the Gaussian autocorrelation here gives unstable oscillating RMSEs up to about $n' = 40$, after which stability is achieved. For the majority of cases, a high value of n' in the stable region resulted in the highest efficiency. In some cases however, as in Figure 3.5, the efficiency in this region was not as high as that obtained by simply taking n' to be twice the number of parameters in the model, for which all methods should agree, being equivalent to equating the first $n'/2$ sample autocorrelation coefficients with their expected values. It should be noted that there is a discrepancy in this respect between use of the Gaussian autocorrelation and use of binary autocorrelation, due to the estimation of the threshold level and the proportion of observations

Model	n	*OLS*		*Spectral*		*WLS*	*GLS*	*Pair*
		(B)	*(G)*	*(B)*	*(G)*			
AR(1)	1000	92	87	**93**	85	88	85	88
	100	**92**	86	**92**	86	87	86	88
MA(1)	100	73	59	**87**	64	59	59	59
ARMA(1,1)	100	46	45	48	**51**	41	49	47
CPU time (seconds)		2	10	6	7	19	6	7

Table 3.2: *Percentage efficiencies relative to MCMC. Each figure is the average over 100 simulations, over a range of parameter values and over the two threshold levels (0 and 1 sd), for optimal n'. The most efficient estimator in each row is highlighted. CPU times are average times for a SunUltra2 to process 100 series for each estimation method, MCMC taking about 10000, 1000, 40000 and 60000 seconds respectively for the four examples shown.*

above the threshold being different from that expected. Further discrepancy can result in practice from some of the parameter estimates hitting the boundary for some of the methods.

3.6.2.1 Summary

We assume the MCMC estimator to be fully efficient and, where results are available, express the efficiency of the other methods relative to this. Table 3.2 shows such percentage efficiencies, the figures shown being the ratio of the average MSE for each method to the average MSE from MCMC. The averages are over different parameter values and threshold levels, for optimal n'. CPU times are averages over the models for all the parameters considered. The MCMC method was too slow to run for MA(1) and ARMA(1,1) processes when $n = 1000$ and so these processes are not included in this table. Table 3.3 shows root mean square errors (RMSEs) for particular examples of each class of model.

From Tables 3.2 and 3.3 we can see that for AR(1) processes, all methods are nearly as efficient as MCMC, and the spectral approach offers no clear benefit over the other fast methods. For MA(1) and ARMA(1,1) processes the spectral method is generally seen to be more efficient than least squares, in some cases the improvement being quite substantial. Weighted and generalised least squares do not appear to offer much advantage over ordinary least squares and appear to be more unpredictable than the other methods. The pairwise likelihood method is fairly stable with n' but generally produces estimates with similar efficiency to ordinary least squares. Note that, for MA(1) processes, the weighted and generalised least squares and pairwise likelihood method are all equivalent to ordinary least squares with $n' = 2$, since the expected values of autocorrelation at lags

Model / Parameter values	n	T	*OLS* (B)	*OLS* (G)	*Spectral* (B)	*Spectral* (G)	*WLS*	*GLS*	*Pair*	*MCMC*
AR(1)	1000	0	**36**	37	**36**	38	**36**	38	**36**	34
$\phi = 0.6$		1	**43**	48	**43**	50	47	50	47	39
AR(1)	100	0	155	155	153	155	153	**145**	154	145
$\phi = 0.3$		1	218	221	218	221	221	**203**	221	206
MA(1)	1000	0	61	61	**54**	**54**	61	61	61	
$\theta = -0.3$		1	70	74	**67**	72	74	74	74	
MA(1)	100	0	291	291	256	**247**	291	291	291	143
$\theta = -0.6$		1	327	339	**321**	333	339	339	339	202
ARMA(1,1)	1000	0	76	76	**70**	73	79	80	75	
$(\phi, \theta) = (0.9, 0.6)$		1	94	95	**91**	96	103	97	93	
ARMA(1,1)	100	0	487	489	451	**311**	488	424	488	275
$(\phi, \theta) = (0.6, -0.3)$		1	621	612	586	**431**	604	539	690	299

Table 3.3: *1000 × RMSE of parameter estimates. Each figure is the average over 100 simulations, for the given parameter values, threshold level T and series length n, at optimal n'. The smallest value in each row is highlighted.*

higher than 1 is zero. Both tables demonstrate that use of the binary autocorrelation is usually to be preferred over the Gaussian. It is interesting and reassuring to note from Table 3.2 the stability in efficiency for AR(1) processes over the two series lengths considered. CPU times for all the fast methods are small enough to be of no consequence. In contrast, times for MCMC are prohibitively high.

3.6.2.2 Discussion

Here, some features of the results of Appendix C are discussed in more detail.

Section C.1 show results for AR(1) processes. RMSEs are seen to generally decrease as ϕ increases. This is intuitive — it seems natural that it would be easier to estimate a parameter in a model with strong correlation than in one with little or no correlation. For series length 1000, RMSEs are roughly half the size for $\phi = 0.9$ as for $\phi = 0.0$ or 0.3. Effects of threshold level and series length are also as might be expected. Series with extreme thresholding clearly contain less information than those thresholded at the mean; this difference is more marked for low values of ϕ. For example, for $\phi = 0.0$ with series length 1000, moving the threshold from 0 to 1 generally increases RMSEs by about 50%, whereas when $\phi = 0.9$ it is more like 20%, although bear in mind that RMSEs are lower anyway for high values of ϕ. Comparing length of series, RMSEs are something like 3–4 times higher for the shorter series, although this varies according to parameter value and threshold level. There are also more problems in finding estimates for the shorter length series, although it only appears to be the spectral method that has problems with estimates being on the boundary, especially for the higher

values of ϕ that are closer to the boundary anyway.

In comparing the methods themselves and the effect of changing n', for OLS the lowest RMSE is typically achieved with a low value of n', e.g. $n' = 2, 4$. The spectral method tends to achieve its best RMSE with higher n' and as n' increases further is usually fairly stable and similar to the lowest value using OLS. The methods using binary autocorrelation nearly always do at least as well or better than the corresponding method using Gaussian autocorrelation. Weighted and generalised least squares do not in general appear to offer much advantage over ordinary least squares, and indeed are often drastically worse. The pairwise likelihood approach appears to be very stable, with RMSEs comparable to the OLS and spectral methods. Comparison with RMSEs from the MCMC show the fast methods to be most efficient for smaller values of ϕ.

Results for MA(1) processes are shown in Section C.2. Given that all autocorrelation coefficients above lag 1 have zero expectation, the only methods to compare are the spectral, with varying n', and OLS with $n' = 2$, which is simply equating the lag 1 correlation to its expected value. These pure MA processes are seen to be rather more problematical than the pure AR processes; we have a lot more cases of estimates being on the boundaries, here involving all methods and for most of the parameter values. We notice that RMSEs increase as θ increases and, as before, that the increase in RMSE due to the higher thresholding is more marked for the parameter values which have the lower RMSEs overall.

The spectral method can be seen to generally behave well for parameter values well away from the boundary, estimates being stable with increasing n' even from a very low value, e.g. $n' = 6$, and with RMSE that is often lower than that from the OLS methods. When the parameters are close to the boundary, even for $\theta = -0.6$, we have problems with estimates falling on the boundary, and this distorts the results, producing instability in the RMSEs.

The MCMC estimator was only considered for MA(1) processes for the shorter length series, due to the large CPU times involved. The fast methods appear relatively more efficient when θ is small. There are anomalies in that e.g. for $\theta = -0.9$, threshold=0 and series length=100, the spectral method using binary autocorrelation produces RMSEs that are lower than from MCMC. This can be attributed to the large numbers of estimates on the boundary (-0.9999) produced here by the spectral method, which when $\theta = -0.9$ have the effect of making the RMSEs artificially low.

Finally, Section C.3 gives results for ARMA(1,1) processes. RMSEs are presented

within each table for ϕ and θ separately. Generally it can be seen that ϕ is estimated with a lower RMSE than θ. Another noticeable feature is the lack of agreement sometimes occurring between the methods for $n' = 4$. It should be the case that all methods produce the same estimates when $n' = 4$, but from inspection of the estimates themselves it can be seen that these discrepancies occur when the optimisation routine is exited with a 'soft-fail'. For example there is quite a large discrepancy between the spectral method and the other methods using Gaussian autocorrelation when $\phi = 0.6$ and $\theta = -0.3$. Inspection of the estimates shows this to be due to the fact that for the spectral method the routine exits a lot of the time with "conditions for a minimum have not all been met but a lower point could not be found", whereas the OLS method manages to find satisfactory minima without fail.

It is hard to quantify the increase in RMSE due to higher threshold or shorter series length for the ARMA(1,1) processes, as the size of the effect is very dependent on the parameter values. This is also true for comparisons over methods and over different values of n', partly due to the effect of unreliable estimates, e.g. on boundary or with a soft-fail message. Again, due to prohibitively high CPU times, the MCMC method has only been considered for series length 100. From the cases shown however, it can be noted that RMSEs from this method are in general considerably lower than those from any of the fast methods.

3.7 Fitting to data

The fitting of ARMA processes to the cow feeding data is now considered, concentrating on the high-protein group of cows. The modelling of visit data and meal data is briefly compared, before going on to look at the effects of allowing for the diurnal pattern of feeding. Implications for the choice of an ARMA(2,1) model are discussed, before presenting parameter estimates and considering goodness of fit.

The data are recorded in continuous time, theoretically accurate to the nearest second. In order to fit an ARMA model we need to transfer this to a discrete time framework. We have chosen to discretise at one minute intervals. This was discussed in Section 1.4.1, and we address the implications for this particular type of model in Section 3.7.2.1.

Cow	$\Phi(\hat{T})$	$\hat{T}$
5	0.9030	1.2988
41	0.9073	1.3243
108	0.8795	1.1735
169	0.9247	1.4374
170	0.9038	1.3035
182	0.9023	1.2948
194	0.9050	1.3106
221	0.8470	1.0237

Table 3.4: *Estimated threshold levels for the eight high-protein cows.* $\Phi(\hat{T})$ *is the proportion of the total time that is non-feeding;* $\hat{T}$ *is the equivalent deviate of the standard normal distribution.*

3.7.1 Autocorrelation structure

Both the binary and Gaussian autocorrelation can be estimated using the methods of Section 3.3. Cows are treated individually, each having minutely data over 30 days resulting in series of length 43200. The threshold levels for the cows are given in Table 3.4.

3.7.1.1 Comparison of visit and meal data

In Allcroft and Glasbey (2000) we chose to suppress short intervals away from feeders and model this derived variable instead, corresponding to feeding bouts or meals i.e. as in Figure 2.11(b). However our modelling approach is also capable of using the visit data directly. Table 3.5 shows autocorrelation estimates up to lag 10, both binary and Gaussian, using each type of data. It can be seen that the difference between the estimates for the two types of data are small, values being marginally higher for the meal data, which is as would be expected as these data are smoother and therefore more correlated then the visit data. Figure 3.6 displays the two types of autocorrelation for each type of data, calculated on a minutely scale for lags up to 6 hours, confirming that there are only negligible differences. Therefore from here on we consider only the modelling of the visit data directly. Our primary aim is the modelling of short-term feeding behaviour and so as we have already argued, it would be desirable to work with models that can allow for behaviour changes at the within-meal level.

Lag	*Visit Data*		*Meal Data*	
	(B)	*(G)*	*(B)*	*(G)*
1	0.8853	0.9885	0.8858	0.9886
2	0.8089	0.9680	0.8113	0.9687
3	0.7433	0.9422	0.7474	0.9439
4	0.6827	0.9115	0.6881	0.9143
5	0.6259	0.8766	0.6326	0.8808
6	0.5733	0.8390	0.5801	0.8438
7	0.5269	0.8014	0.5340	0.8070
8	0.4850	0.7640	0.4926	0.7706
9	0.4496	0.7296	0.4576	0.7371
10	0.4133	0.6916	0.4215	0.6999

Table 3.5: *Estimates of autocorrelation for Cow 108 to show similarity between values based on visit data and meal data, for both binary (B) and Gaussian (G) autocorrelation.*

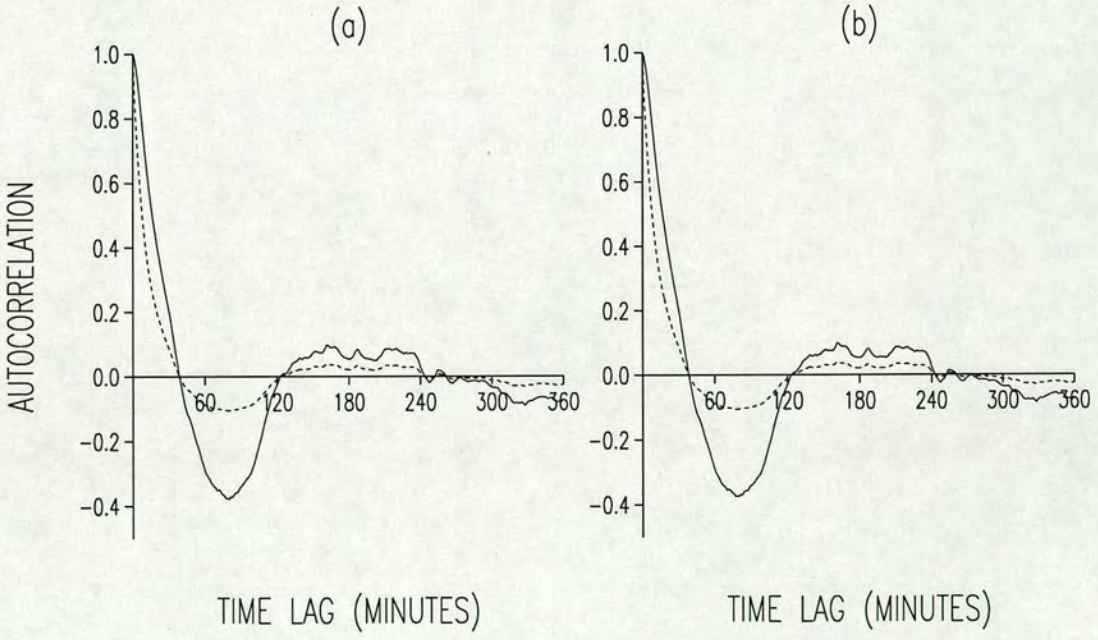

Figure 3.6: *Autocorrelation estimates for Cow 108; (a) visit data, (b) meal data; (- - -) binary autocorrelation, $\hat{\rho}^{(B)}$; (—) Gaussian autocorrelation, $\hat{\rho}^{(G)}$.*

3.7.1.2 Allowing for diurnal trend

As discussed in Section 3.3.3, the diurnal trend can be taken into account by generalising the constant threshold T to one that varies with time, denoted T_t. In theory it would be desirable to do this parametrically, e.g. sinusoidally. This might be suitable for a cow with only a weak diurnal cycle, e.g. Cow 5, but for a cow with a strong diurnal cycle such as Cow 108, the pattern would be hard to capture without including a large number of harmonics. The approach we take therefore is to calculate, on a minutely basis, the proportion of days for which feeding occurred during that minute of the day. Working on this minutely scale results in a very non-smooth estimate of the trend, therefore we smooth using a moving average.

Letting p_t be the overall minutely probability of feeding in the t-th minute of the day, for $t = 1, \ldots, 1440$, based on the average over the 30 days, we use $\tilde{p}_t^{(L)}$ to denote the moving average of length $2L + 1$, given by

$$\tilde{p}_t^{(L)} = \frac{1}{2L+1} \sum_{l=-L}^{L} p_{t+l}.$$

Cross-validation was used to find the optimal length for the moving average, the approach being to take the 30 days of data for each cow, omit each day in turn, and calculate a Bernoulli likelihood of the form

$$\sum_{r=1}^{30} \sum_{t=1}^{1440} x_{rt} \log \tilde{p}_{rt}^{(L)} + (1 - x_{rt}) \log \left(1 - \tilde{p}_{rt}^{(L)}\right),$$

where x_{rt} are the feeding data (0 if not feeding, 1 if feeding) for the t-th minute of day r, and $\tilde{p}_{rt}^{(L)}$ is the estimated probability of feeding in the tth minute of the day estimated from 29 days of data, i.e. with day r omitted, using a moving average of length $2L + 1$. A range of L was considered, from 0, increasing in 5 minute intervals, up to 120 minutes. Figure 3.7 shows how the log-likelihood changes as L changes for Cows 5 and 108, and Table 3.6 shows the resulting optimal moving average lengths for all eight high-protein cows. Note that the optimal length for animals with only a weak diurnal cycle is larger than for those displaying a strong cycle. This is intuitive. It is also true that the likelihood is flatter around the maximum for those animals with less diurnal pattern. This is also intuitive. Figures 3.8 and 3.9 show the resulting estimates of diurnal pattern for Cows 5 and 108, plotted on the same scale. As already remarked, Cow 5 shows little diurnal pattern whereas Cow 108 shows a much stronger pattern.

Table 3.7 shows autocorrelation estimates up to lag 10 for Cows 5 and 108, (B) indicating binary, (G) Gaussian, both ignoring the diurnal pattern, and (T) the

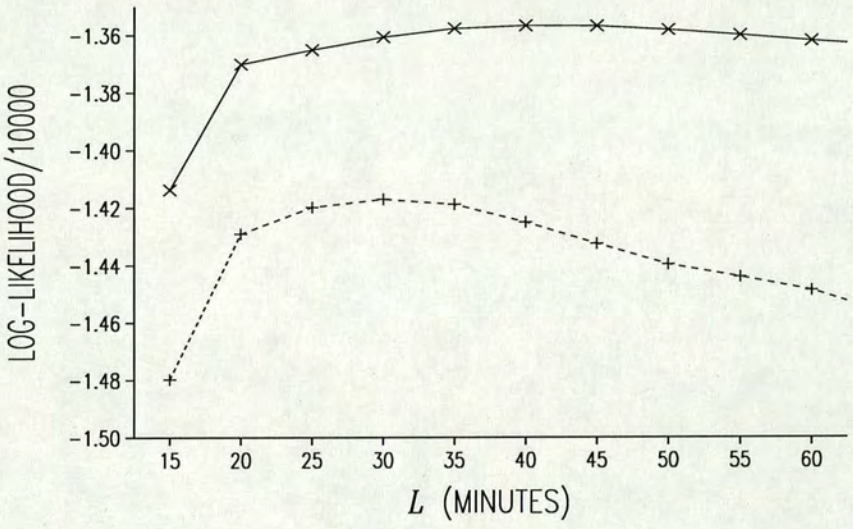

Figure 3.7: *Bernoulli-type log-likelihood used in cross-validation to determine optimal length for moving average smoothing;* (×—×) *Cow 5,* (+– – –+) *Cow 108.*

Cow	*L*
5	40
41	45
108	30
169	55
170	35
182	30
194	75
221	25

Table 3.6: *Optimal values of L in minutes, corresponding to a moving average of length* $(2L + 1)$ *minutes.*

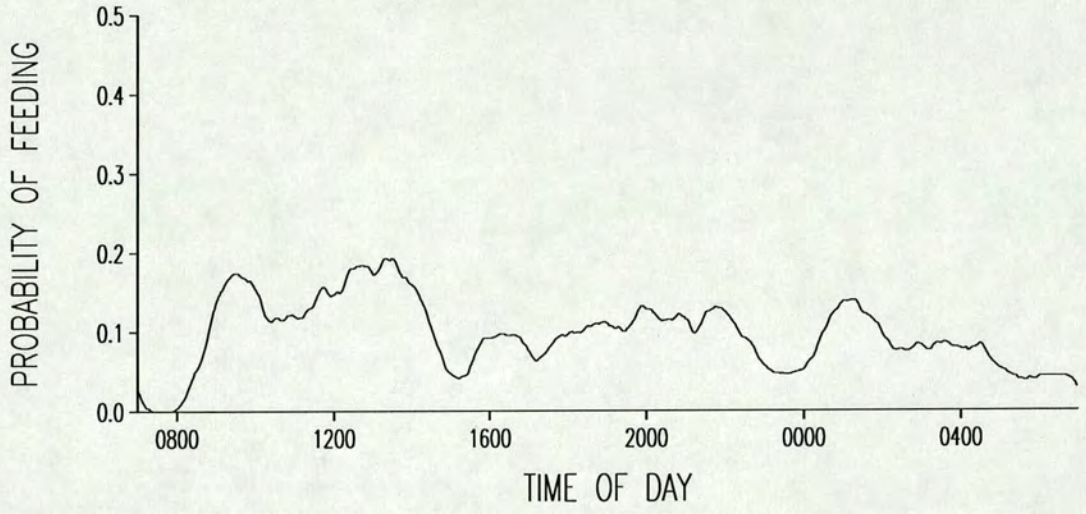

Figure 3.8: *Probability of feeding over the day for Cow 5, a cow with a weak diurnal cycle. Probabilities are calculated from minutely estimates using a moving average of length 81 minutes ($L = 40$).*

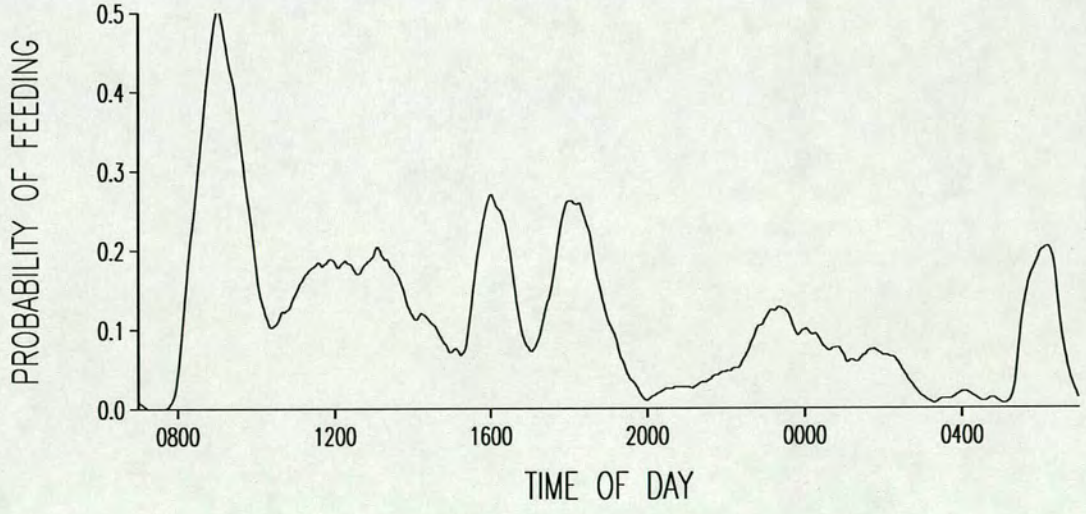

Figure 3.9: *Probability of feeding over the day for Cow 108, a cow with a strong diurnal cycle. Probabilities are calculated from minutely estimates using a moving average of length 61 minutes ($L = 30$).*

Lag	*Cow 5*			*Cow 108*		
	(B)	*(G)*	*(T)*	*(B)*	*(G)*	*(T)*
1	0.8853	0.9892	0.9880	0.8853	0.9885	0.9845
2	0.8033	0.9682	0.9645	0.8089	0.9680	0.9577
3	0.7411	0.9447	0.9383	0.7433	0.9422	0.9246
4	0.6849	0.9179	0.9085	0.6827	0.9115	0.8853
5	0.6398	0.8924	0.8801	0.6259	0.8766	0.8408
6	0.5950	0.8634	0.8479	0.5733	0.8390	0.7920
7	0.5536	0.8334	0.8147	0.5269	0.8014	0.7437
8	0.5156	0.8030	0.7812	0.4850	0.7640	0.6963
9	0.4803	0.7723	0.7474	0.4496	0.7296	0.6533
10	0.4455	0.7395	0.7115	0.4133	0.6916	0.6057

Table 3.7: *Estimates of autocorrelation for Cows 5 and 108; binary (B), Gaussian ignoring trend (G) and Gaussian allowing for diurnal trend (T).*

Gaussian taking the diurnal effect into account. Comparison of the Gaussian autocorrelation ignoring and allowing for trend shows the estimates to be lower when trend has been allowed for. This is as we would expect; allowing for the trend has the effect of removing some of the correlation.

3.7.2 Choice of model

Inspection of the autocorrelation structure for all cows shows the general pattern to be as in Figure 3.6 which shows $\hat{\rho}^{(B)}$ and $\hat{\rho}^{(G)}$ for Cow 108. The simplest class of ARMA model that can produce this shape is ARMA(2,1), the autocorrelation function for which is a mixture of two exponential functions. At this point we address issues concerning the arbitrary minutely time scale that we have chosen to work with. It is important that the fitted model is invariant to the discretisation scale chosen, and additionally we would like the model to have a continuous time analogue, since in theory the feeding data are recorded to the accuracy of seconds. We will see in the following sections that the ARMA(2,1) process has most of these desired properties.

3.7.2.1 Discretisation of time

Although usually considered in a discrete time framework, ARMA processes can also occur in continuous time. However not all classes of ARMA model in discrete time translate into continuous time. We briefly consider here which classes of ARMA models remain in the same class of model when the time unit is changed.

Consider the AR(2) process,

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} = \epsilon_t,$$

where terms are as described for the ARMA(2,1) process in Section 3.2.4.

The autocorrelation at lag l for this process can be written in the form

$$\rho_l = a\pi_1^l + (1-a)\pi_2^l, \tag{3.16}$$

where a, π_1 and π_2 are determined by the parameters ϕ_1 and ϕ_2.

If we then consider doubling the time-step, the autocorrelation at integer lags is now the set of even-lag autocorrelations from the original set. If this is still to be described as an AR(2) process then its autocorrelation will be of the same form, i.e.

$$\rho_l = b\omega_1^l + (1-b)\omega_2^l$$

for some b, ω_1 and ω_2.

For these to match what we had previously and describe the same process, we need to satisfy

$$\begin{aligned} \omega_1 &= \pi_1^2 \\ \omega_2 &= \pi_2^2 \\ \text{and} \quad a &= b. \end{aligned}$$

These are all functions of ϕ_1 and ϕ_2, hence we have three equations in two unknowns, and in general these cannot be satisfied. If however we introduce an extra parameter by generalising to an ARMA(2,1) process, then in principle we have the extra parameter needed to solve the equations. Hence since the ARMA(2,1) process, $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t - \theta_1 \epsilon_{t-1}$, has autocorrelation of the same form as (3.16) it does possess the property that the time unit can be changed and the process remains in the same class of process, with new parameters related to the original. Note though that there may be restrictions on parameter values, for example take an AR(1) process with $\phi < 0$. The autocorrelation at integer lags has alternating sign, therefore it does not make sense to double or halve the time-step.

For the general ARMA(p,q) process, given by

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q},$$

the autocorrelation function is given by the solution of

$$\rho_l - \phi_1 \rho_{l-1} - \ldots - \phi_p \rho_{l-p} = 0 \quad \text{for } l \geq \max(p, q+1).$$

The solution to this is of the form

$$\rho_l = a_1\pi_1^l + \ldots + a_p\pi_p^l.$$

for $l \geq q+1-p$, For this this to hold for all lags ≥ 0 we must consider processes for which $p \geq q+1$, giving an autocorrelation function of the form of a mixture of exponential functions (where the π_i are real) or damped sine waves (where the π_i occur in complex pairs), see for example Box et al. (1994, page 79).

So in general, an ARMA(2,1) model will satisfy the condition of being invariant with respect to timescale — the only limiting factor is that as the timescale is decreased linearly, the parameters get geometrically closer to the boundaries of their parameter-space. This can cause problems when parameters are being estimated using numerical methods. Therefore we carry out our model fitting using a minutely timescale, which is sufficiently detailed to capture the main features of the data, whilst not causing too many problems with parameters being near the boundaries.

3.7.2.2 Continuous time

For an ARMA model to have a continuous-time analogue which is locally smooth, it must have a continuous autocorrelation function (Chatfield, 1996, page 78, section 3.4.8). In particular this means that the first derivative must be zero at $l = 0$, or equivalently the second derivative must exist at 0. Otherwise we will have a cusp or a discontinuity at the origin. The ARMA(2,1) model has such a discontinuity at the origin and therefore consideration of this process in continuous time has problems. Cox and Miller (1965, Section 7.4) discuss this in detail with reference to Ornstein-Uhlenbeck diffusion processes, and show that for a process with an autocorrelation function not differentiable twice at zero, realisations have short term jitter. Hence if the process crosses the threshold at a time t_0, then for any small $\epsilon > 0$, the process will cross the threshold infinitely often in the interval $(t_0, t_0 + \epsilon)$. Lindgren and Rychlik (1991) also consider this problem and use the theory of Slepian models to derive approximations for expressions such as the distribution of lengths of excursions above the threshold. These again depend on the existence of the second derivative at zero, and hence cannot be applied here. Therefore, although the ARMA(2,1) model fit in discrete time will still be valid for ever smaller discretisation scales, in the limit as the time step tends to zero, there are problems.

3.7.3 Model estimation

We will estimate parameters for the ARMA(2,1) model fit on a minutely time scale and then, since there are no analytic forms for the variation associated with the binary and tetrachoric correlations, we use simulation to check whether the data are consistent with the fitted model. To do this we simulate series with the parameters as estimated and see if the data fall within the range of sampling variation.

3.7.3.1 Comparison of estimation methods via simulation

The simulation study of Section 3.6 did not consider ARMA(2,1) models so we first briefly consider simulation of series with parameters close to their estimated values. Series length was the same as the data, i.e. 43200, and parameter values were as estimated for Cow 108 and with a threshold level to match the observed feeding rate for this cow.

For the simulated series we estimate parameters as previously using the minimisation routine E04JAF in the NAG library (Numerical Algorithms Group, 1993) using Fortran 90. This routine only allows simple bounds on the parameters, so in order to satisfy the linear constraints for (ϕ_1, ϕ_2) as described in Section 3.2.4, we modify the set of parameters to be estimated from $\{\phi_1, \phi_2, \theta\}$ to $\{\phi_1, 2\phi_2/(1-|\phi_1|), \theta\}$. Now simple bounds can be imposed, constraining each to lie in the interval $(-1, 1)$.

Table 3.8 shows results for the least squares and spectral estimators only, using both binary and Gaussian autocorrelation. Root mean square errors here are the square root of the sum of the three individual mean square errors for ϕ_1, ϕ_2 and θ. Least squares using the binary autocorrelation is seen to be the only method that is fully efficient with a low value of n', whilst the other methods stabilise as n' is increased. For small n' there are problems with the spectral method using either type of autocorrelation and for least squares using the Gaussian autocorrelation. This is mainly due to the distortion of the results when for some series parameters are estimated on the boundaries. Results are further distorted when the three individual parameter MSEs are combined. Nevertheless it is reassuring that all methods perform well for $n' > 240$. For series of length 43200, this is still a relatively small number of lags to consider. It was hoped that the sizes of the autocorrelation and inverse autocorrelation coefficients at this point would have an order of magnitude of around 10^{-4}, to fit in with the broad conclusions

n'	*OLS*		*Spectral*	
	(B)	*(G)*	*(B)*	*(G)*
6	5	1294	1321	651
10	5	594	829	356
20	5	220	284	126
60	5	39	36	281
120	5	15	24	21
240	5	7	7	9
360	5	7	7	9
720	5	7	7	6

Table 3.8: *1000 × joint RMSEs of parameter estimates for simulations of an ARMA(2,1) model. Each figure is the average over 100 simulations for series of length 43200 minutes (30 days), with parameter values* $\phi = (1.9716, -0.9728), \theta = -0.9927$ *and threshold level* $T = 1.1735$*, as estimated for Cow 108.*

made at the end of Section 3.4.5. But here, for $l > 120$ we have autocorrelation $\rho_l < 0.1$ and $\alpha_l < 2.5 \times 10^{-5}$. So the values of the inverse autocorrelation are as expected, but those of the autocorrelation have not decayed sufficiently by this point. However earlier, and in the simulation study, only AR(1), MA(1) and ARMA(1,1) processes were considered and so more work needs to be done in order to draw general conclusions about the size of n' required for higher-order ARMA processes.

3.7.3.2 Parameter estimates

Table 3.9 shows parameter estimates for Cows 5, 41 and 108, obtained with the least squares method and using either binary or Gaussian autocorrelation and also allowing for trend. The values of n' shown are those above which the parameter estimates become stable and do not change as n' is further increased. In most cases taking n' substantially lower than this has little effect on estimates.

Figures 3.10 and 3.11 show the estimated Gaussian autocorrelation for Cows 41 and 108, along with 95% simulation-based pointwise confidence intervals which were obtained as the 5th and 195th estimates of the ordered sample autocorrelations from 199 simulated series. For Cow 41 the model simulated from was that estimated by least squares using the Gaussian autocorrelation allowing for trend, i.e. $\phi = (1.9771, -0.9777), \theta = -1.000$; for Cow 108 we used the model ignoring trend, $\phi = (1.9716, -0.9728), \theta = -0.9927$, see Table 3.9. The fit for Cow 108 is seen to be good, the estimated autocorrelation being consistent with the model for the majority of the time. For Cow 41 the fit is not quite as good, but still reasonable. Similar pictures were obtained for the other cows.

Cow		n'	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\theta}$
5	*(B)*	720	1.9495	−0.9504	−0.9956
5	*(G)*	960	1.9695	−0.9704	−1.0000
5	*(T)*	960	1.9697	−0.9705	−1.0000
41	*(B)*	960	1.9554	−0.9559	−0.9979
41	*(G)*	1440	1.9768	−0.9773	−1.0000
41	*(T)*	1440	1.9771	−0.9777	−1.0000
108	*(B)*	720	1.9526	−0.9541	−0.9808
108	*(G)*	960	1.9716	−0.9728	−0.9927
108	*(T)*	720	1.9688	−0.9700	−1.0000

Table 3.9: *Parameter estimates for ARMA(2,1) models for Cows 5, 41 and 108, using least squares with binary (B) correlations, Gaussian correlations ignoring trend (G) and Gaussian correlations allowing for trend (T). Also shown are the values of n' above which the parameter estimates are stable.*

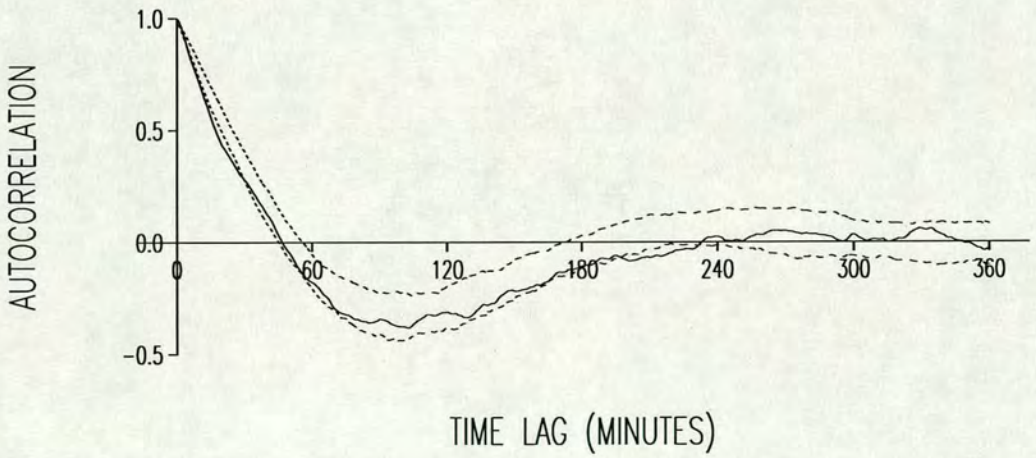

Figure 3.10: *Gaussian autocorrelation for Cow 41; (—) $\hat{\rho}^{(G)}$, (- - -) simulation-based 95% confidence envelope for $\hat{\rho}^{(G)}$ for the fitted ARMA(2,1) model with parameters $\phi = (1.9771, -0.9777), \theta = -1.000$.*

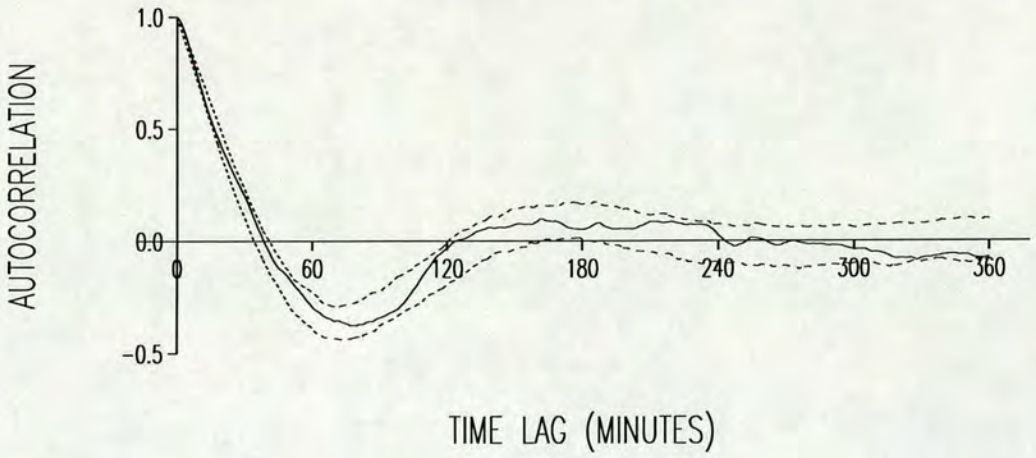

Figure 3.11: *Gaussian autocorrelation for Cow 108; (—) $\hat{\rho}^{(G)}$, (- - -) simulation-based 95% confidence envelope for $\hat{\rho}^{(G)}$ for the fitted ARMA(2,1) model with parameters $\phi = (1.9716, -0.9728), \theta = -0.9927$.*

3.7.3.3 Model validation

Figure 3.1 showed a sample of data simulated from the fitted model for Cow 108, together with the corresponding realisation of the underlying latent variable. The feeding patterns are not dissimilar to the observed data of Appendix A. To check the model further, Figures 3.12 and 3.13 show the marginal distributions of feeding durations and non-feeding durations, comparing the distributions from the data with those obtained by simulation of the fitted model, since there are no closed forms for these distributions. The form of these distributions played no part in the motivation for this model, so it is reassuring to observe that the marginal distributions are of the correct shape, in particular capturing the bimodality of the distribution of durations between feeder-visits.

3.8 Summary

Assuming the observed binary data to have arisen from the thresholding of an underlying continuous variable is a biologically plausible model for the cow feeding data. This has been the motivation for an investigation into parameter estimation for censored ARMA processes. Firstly it is important to note the one-to-one correspondence between the autocorrelation structure of the observed binary process and that of the unobserved underlying Gaussian process. I discussed the use of a tetrachoric method to estimate the latter, before generalising to make allowance

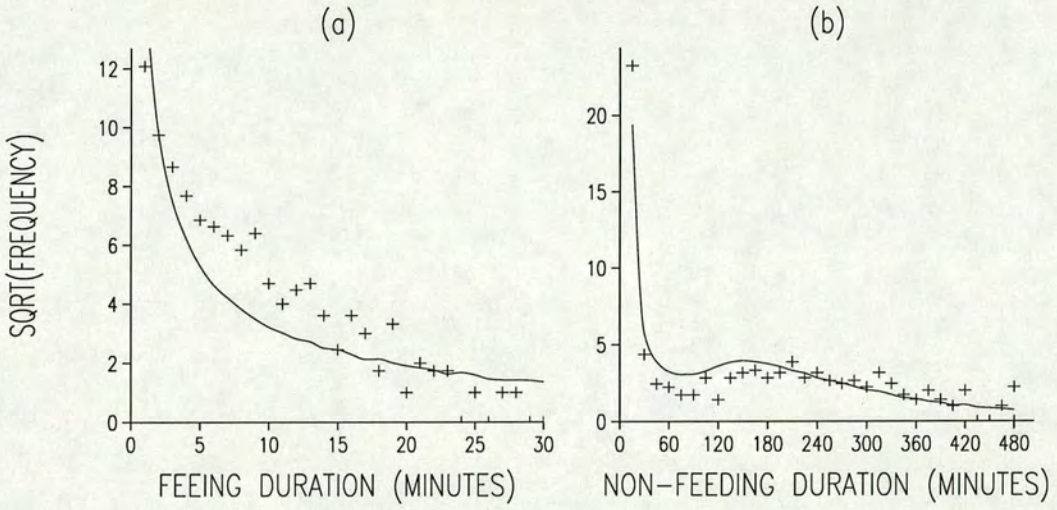

Figure 3.12: *Cow 41; marginal distributions of (a) feeding durations, (b) non-feeding durations; (+) sample frequencies, (—) frequencies based on 199 realisations of the fitted ARMA(2,1) model.*

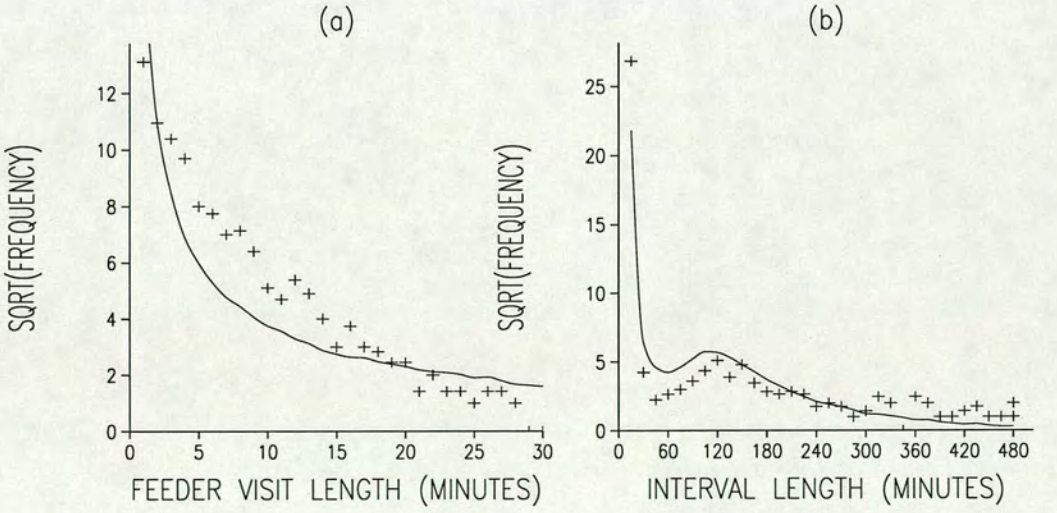

Figure 3.13: *Cow 108; marginal distributions of (a) feeding durations, (b) non-feeding durations; (+) sample frequencies, (—) frequencies based on 199 realisations of the fitted ARMA(2,1) model.*

for the diurnal pattern present in the data. I then went on to consider methods of parameter estimation, comparing computationally fast procedures with a method using MCMC that was considered to be fully efficient. It was seen that a method based on maximisation of a spectral quasi-likelihood was computationally quick and sometimes more efficient than other fast methods based on least squares. As ARMA processes have only short-term memory, I investigated the number of lags needing to be retained for model estimation, finding it to be much less than the full set available and dependent on the speed of decay of the autocorrelation and inverse autocorrelation coefficients. This greatly reduces the computational burden of having to consider the whole set of autocorrelation coefficients. The main advantage of the spectral method over the other methods is that the exact choice of how many lags to retain is not critical. Simulations showed that using the binary autocorrelation directly was preferable to the Gaussian, both in terms of efficiency and stability. For the ARMA(2,1) model, the least squares method using binary correlations appears to be the only one stable from very low n', but the other methods are efficient once n' is taken large enough. In fitting to the data, the fit of the models was found to be acceptable, and on inspection of the marginal distributions of feeding and non-feeding durations produced by the model, these were found to be consistent with the data. The latent Gaussian model therefore appears to be a flexible one for categorical behaviour data and also has scope for extension to a multivariate framework and inclusion of other covariates.

Chapter 4

The spectral representation of the likelihood for a stationary Gaussian time series

In the previous chapter I used the spectral representation of the likelihood, in its restricted form, as a method of estimating parameters for ARMA processes. In Section 4.1, the background to this is outlined and, for completeness, some results are stated that will be used for the proofs. In Section 4.2, the univariate case is dealt with — Section 4.2.1 introduces the notation and states the results to be proved, Section 4.2.2 gives a clear derivation for the full spectral likelihood, and Section 4.2.3 proves how this is well-approximated by the restricted form of the likelihood. Section 4.3 contains corresponding sections for the multivariate case. This includes the univariate case as a special case, but if the univariate case is all that is required, as is the case when we used the method in Chapter 3, it is simpler to consider Section 4.2 only. A brief outline of this proof is contained in Allcroft and Glasbey (2000), with a more detailed version in the appendix of Allcroft and Glasbey (2001).

4.1 Background

The representation of the likelihood of a stochastic process in its spectral form rather than directly was first suggested by Whittle (1953). For a multivariate stationary time series he maximised it in order to provide the maximum Gaussian likelihood parameter estimates, i.e. the maximum likelihood estimates assuming a Gaussian model. The method was particularly useful with the limited computing power then available, but is not used widely these days, due to the availability of

recursive techniques based on the Kalman filter and state-space representations, allowing exact Gaussian likelihoods to be computed efficiently (see for example Lütkepohl, 1991). However, for censored data, state-space representations do not exist which is why we revisit the spectral representation.

Whittle presents a proof for a multivariate series, corresponding to Section 4.3.2 here, although he uses the continuous Fourier transform and we present the proof in terms of the discrete Fourier transform, as used by e.g. Chandler (1996). The result also appears in Brillinger (1975, page 238) and Glasbey et al. (1998), both without proof, which is why we present a detailed derivation here.

It should be noted that the spectral form is an exact expression in the case for which a circular covariance structure is assumed, i.e. when $\gamma_l = \gamma_{n-l}$, where n is the series length and l the lag, for $l = 0, 1, \ldots, n$. Otherwise the approximation is asymptotic as $n \to \infty$. Some formal results about the goodness of the approximation are given in Coursol and Dacunha-Castelle (1983).

For short-term memory processes, e.g. ARMA processes, for which the autocorrelation decays exponentially, most of the useful information about the process is contained in the first few lags and the high-lag sample autocorrelation coefficients are mostly noise. Therefore it would seem logical to discard the higher lag coefficients and just make use of those at short lags. We prove that for short-term memory processes, instead of taking the full likelihood we get a good approximation by replacing the Fourier transforms of the full set of auto- and cross-covariances by only the first few lags.

4.1.1 The trapezoidal rule for integration

In Sections 4.2.3 and 4.3.3 we use a result related to the trapezoidal rule for integration which is stated here, as in Burden and Faires (1985, page 165).

Let $h = (b - a)/n$ and $x_j = a + jh$ for $j = 0, 1, \ldots, n$. Then for a continuous function $f(x)$ defined on the interval $[a, b]$, the composite trapezoidal rule for integration approximates the integral

$$\int_{x=a}^{b} f(x)dx$$

by

$$\frac{h}{2}\left[f(a) + f(b) + 2\sum_{j=1}^{n-1} f(x_j)\right] \tag{4.1}$$

and the error in this approximation is given by

$$-\frac{(b-a)h^2}{12}f''(\eta) \tag{4.2}$$

for some $\eta \in (a, b)$.

4.2 Univariate series

4.2.1 Notation

For a stationary Gaussian mean-corrected time series of length n, i.e. $y = (y_0, y_1, \ldots, y_{n-1})$, the spectral representation of the log-likelihood, $\mathcal{L}$, is given by

$$\mathcal{L} = -\frac{1}{2}\sum_{k=0}^{n-1}\log S_k - \frac{1}{2}\sum_{k=0}^{n-1}\frac{\hat{S}_k}{S_k}, \tag{4.3}$$

where S_k is the spectrum and $\hat{S}_k$ the periodogram of y.

We can define the spectrum S_k, which is real, as the discrete Fourier transform of the theoretical autocovariance coefficients γ_l, and the periodogram $\hat{S}_k$ as the discrete Fourier transform of the sample autocovariances $\hat{\gamma}_l$. At the kth canonical frequency, $2\pi k/n$, these are given by

$$S_k = \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1}\gamma_l e^{\frac{-2\pi i k l}{n}} \quad \text{for } k = 0, 1, \ldots, n-1, \tag{4.4}$$

and

$$\hat{S}_k = \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1}\hat{\gamma}_l e^{\frac{-2\pi i k l}{n}} \quad \text{for } k = 0, 1, \ldots, n-1, \tag{4.5}$$

where $i = \sqrt{-1}$ and the sample autocovariance coefficient at lag l is defined as

$$\hat{\gamma}_l = \frac{1}{n}\sum_{j=0}^{n-1} y_j y_{j+l \bmod n}. \tag{4.6}$$

Throughout, for convenience, we assume n is even. If n is odd then the limits on the sums in (4.4) and (4.5) change to $\pm(n-1)/2$ and subsequent expressions change accordingly.

Note that we are assuming a circular model for the autocovariance structure, i.e. that $\gamma_l = \gamma_{n-l}$ (see Section 3.3.1) and so (4.3) can also be written as

$$\mathcal{L} = -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\log S_k - \frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\frac{\hat{S}_k}{S_k}, \tag{4.7}$$

as the sums are circular. The original expression (4.3) is more intuitive, but the second (4.7) is more useful when we consider the restricted likelihood below. If (4.7) is used, then (4.4) and (4.5) can be equivalently defined for $k = -n/2, \ldots, n/2 - 1$. We now define the *restricted* form of the log-likelihood $\mathcal{L}'$ by

$$\mathcal{L}' = -\frac{n}{2n'} \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log S'_k - \frac{n}{2n'} \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \frac{\hat{S}'_k}{S'_k}, \tag{4.8}$$

where $n' < n$.

S'_k is the *restricted spectrum*, i.e. the Fourier transform of the autocovariances up to lag $n'/2$ only, i.e.

$$S'_k = \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} \gamma_l e^{\frac{-2\pi i k l}{n'}} \quad \text{for} \quad k = -\frac{n'}{2}, \ldots, \frac{n'}{2} - 1, \tag{4.9}$$

and similarly we have the *restricted periodogram*,

$$\hat{S}'_k = \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} \hat{\gamma}_l e^{\frac{-2\pi i k l}{n'}} \quad \text{for} \quad k = -\frac{n'}{2}, \ldots, \frac{n'}{2} - 1.$$

We will show in Section 4.2.3 that for short-term memory processes such as ARMA models, $\mathcal{L}$ can be approximated by $\mathcal{L}'$, and therefore for some $n' < n$, maximisation of $\mathcal{L}'$ results in parameter estimates that are good approximations to those resulting from the maximisation of $\mathcal{L}$ (Glasbey et al., 1998).

4.2.1.1 Inverse autocovariances

The inverse autocovariance coefficients of a time series are defined as the autocovariance coefficients associated with the inverse of the spectral density of the series. They were introduced by Cleveland (1972) and further discussed by Chatfield (1979). For a univariate series, the autocovariance coefficient at lag l can be written as the inverse Fourier transform of the spectrum, given by (4.4), i.e.

$$\gamma_l = \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} S_k e^{\frac{2\pi i k l}{n}} \quad \text{for} \quad l = 0, 1, \ldots, n-1.$$

The *inverse autocovariance coefficient* at lag l can be written similarly as the inverse Fourier transform of the inverse of the spectrum, i.e.

$$\alpha_l = \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \frac{1}{S_k} e^{\frac{2\pi i k l}{n}} \quad \text{for} \quad l = 0, 1, \ldots, n-1. \tag{4.10}$$

Again, we could alternatively define these quantities for $l = -n/2, \ldots, n/2 - 1$.

We also define the *restricted inverse autocovariance coefficient* at lag l as

$$\alpha'_l = \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \frac{1}{S'_k} e^{\frac{2\pi i k l}{n'}} \quad \text{for } l = -\frac{n'}{2}, \ldots, \frac{n'}{2} - 1, \tag{4.11}$$

i.e. the inverse Fourier transform of the inverse of the restricted spectrum.

Chatfield (1979) notes the interesting fact that if we take the ARMA model given by

$$\Phi(B)X_t = \Theta(B)\epsilon_t,$$

then the inverse autocovariance coefficients are the autocovariance coefficients of the inverse process

$$\Theta(B)X_t = \Phi(B)\epsilon_t.$$

Here, X_t is the current value of the process and the ϵ_t are independent normally distributed observations. B is the backshift operator, $\Phi(\cdot)$ denotes the autoregressive part of the model and $\Theta(\cdot)$ the moving average part.

Finally, it is worth noting that the Fourier transform of a real series is Hermitian. If, in addition, the real series is circular, as defined in Section 3.3.1, then the resulting Fourier transform is real. Also note that the (inverse-) Fourier transform of a Hermitian series is real.

4.2.2 Full likelihood

Here we provide a derivation of the spectral representation of the log-likelihood for a univariate series, given by (4.3).

The log-likelihood $\mathcal{L}$ for the series y is given by

$$\mathcal{L} = -\frac{1}{2}\log|V| - \frac{1}{2}y^T V^{-1} y, \tag{4.12}$$

where V is the variance matrix of the series, with elements $V_{jk} = \gamma_{|j-k|}$.

We introduce an $n \times n$ matrix F with elements $F_{jk} = e^{\frac{-2\pi i j k}{n}}/\sqrt{n}$, with $j, k = 0, \ldots, n-1$, the inverse of which is its complex conjugate, $\bar{F}$, because

$$\begin{aligned}(F\bar{F})_{jk} = (\bar{F}F)_{kj} &= \frac{1}{n}\sum_{l=0}^{n-1} e^{\frac{-2\pi i (j-k) l}{n}} \\ &= \frac{1}{n} \times n\delta_{jk} = \delta_{jk},\end{aligned}$$

where

$$\delta_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases}$$

Using standard results from matrix algebra we can rewrite (4.12) as

$$\mathcal{L} = -\frac{1}{2}\log|\bar{F}(FV\bar{F})F| - \frac{1}{2}(\bar{F}y)^T(FV\bar{F})^{-1}Fy, \tag{4.13}$$

and we show that (4.3) is equivalent to this.

First we show that $(FV\bar{F})$ is diagonal, with kth diagonal element S_k. We have

$$\begin{aligned}
(FV\bar{F})_{jk} &= \frac{1}{n}\sum_{r=0}^{n-1}\sum_{m=0}^{n-1} e^{\frac{-2\pi ijm}{n}}\gamma_{|m-r|}e^{\frac{2\pi irk}{n}} \\
&= \frac{1}{n}\sum_{r=0}^{n-1} e^{\frac{-2\pi i(j-k)r}{n}}\sum_{l=-r}^{n-r-1}\gamma_{|l|}e^{\frac{-2\pi ijl}{n}}, \qquad (\text{put } l = m-r) \\
&= \frac{1}{n}\left(\sum_{r=0}^{n-1} e^{\frac{-2\pi i(j-k)r}{n}}\right)\left(\sum_{l=0}^{n-1}\gamma_l e^{\frac{-2\pi ijl}{n}}\right),
\end{aligned}$$

since the sums are circular, and so the indices on the second sum can be made independent of r. The first sum is now equal to $n\delta_{jk}$ and the second is equivalent to the spectrum S_j as given by (4.4). Therefore

$$(FV\bar{F})_{jk} = S_j\delta_{jk},$$

and for the first term on the right hand side of (4.13) we have

$$\begin{aligned}
\log|\bar{F}(FV\bar{F})F| &= \log\left(|\bar{F}F||FV\bar{F}|\right) \\
&= \log\left(1 \times \prod_{k=0}^{n-1} S_k\right) \\
&= \sum_{k=0}^{n-1}\log S_k \qquad \text{as required.}
\end{aligned}$$

The discrete Fourier transform of the data series y is given by

$$f_k = \frac{1}{\sqrt{n}}\sum_{j=0}^{n-1} y_j e^{\frac{-2\pi ijk}{n}} \qquad \text{for } k = 0, 1, \ldots, n-1,$$

and we have that

$$(Fy)_k = \sum_{j=0}^{n-1}\frac{1}{\sqrt{n}}e^{\frac{-2\pi ijk}{n}}y_j = f_k,$$

and similarly

$$(\bar{F}y)_k = \bar{f}_k.$$

Also

$$
\begin{aligned}
|f_k|^2 = f_k \bar{f}_k &= \left(\frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} y_m e^{\frac{-2\pi i m k}{n}}\right)\left(\frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} y_j e^{\frac{+2\pi i j k}{n}}\right) \\
&= \frac{1}{n} \sum_{m=0}^{n-1} \sum_{j=0}^{n-1} y_m y_j e^{\frac{-2\pi i (m-j) k}{n}} \\
&= \frac{1}{n} \sum_{j=0}^{n-1} \sum_{l=-j}^{n-1-j} y_j y_{j+l} e^{\frac{-2\pi i l k}{n}} \qquad (\text{put } l = m - j).
\end{aligned}
$$

Again the sums are circular and so we can rewrite the second one independently of j. Doing this, changing the order of summation and using (4.6) and (4.5) we have

$$
\begin{aligned}
|f_k|^2 &= \sum_{l=0}^{n-1} e^{\frac{-2\pi i k l}{n}} \left(\frac{1}{n} \sum_{j=0}^{n-1} y_j y_{j+l}\right) \\
&= \sum_{l=0}^{n-1} \hat{\gamma}_l e^{\frac{-2\pi i k l}{n}} \\
&= \hat{S}_k.
\end{aligned}
$$

So the second term on the right hand side of (4.13) is

$$
\begin{aligned}
(\bar{F}y)^T (FV\bar{F})^{-1} Fy &= \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} \bar{f}_j \frac{1}{S_j} \delta_{jk} f_k \\
&= \sum_{k=0}^{n-1} \frac{|f_k|^2}{S_k} \\
&= \sum_{k=0}^{n-1} \frac{\hat{S}_k}{S_k} \qquad \text{as required.}
\end{aligned}
$$

Therefore (4.3) is equivalent to (4.13) and therefore also (4.12), providing the proof of (4.3) or, equivalently, of (4.7).

4.2.3 Restricted likelihood

Here we provide a proof that the restricted log-likelihood, $\mathcal{L}'$, given by (4.8), is an approximation for the full log-likelihood, $\mathcal{L}$, given by (4.7).

Using the periodogram definition (4.5), the full log-likelihood (4.7) can be written

$$
\begin{aligned}
\mathcal{L} &= -\frac{1}{2} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log S_k - \frac{1}{2} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \frac{\sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \hat{\gamma}_l e^{\frac{-2\pi i k l}{n}}}{S_k} \\
&= -\frac{1}{2} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log S_k - \frac{1}{2} \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \left[\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \frac{1}{S_k} e^{\frac{-2\pi i k l}{n}}\right] \hat{\gamma}_l.
\end{aligned}
$$

The inverse spectrum is real and circular, hence the result of the Fourier transform in the square bracket is real and this term is equivalent to the definition of the inverse autocovariance coefficient at lag l (4.10) and we can write

$$\mathcal{L} = -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log S_k - \frac{1}{2}\sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \alpha_l \hat{\gamma}_l. \tag{4.14}$$

Similarly for $\mathcal{L}'$ (4.8) we can write

$$\mathcal{L}' = -\frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log S'_k - \frac{n}{2n'}\sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} \alpha'_l \hat{\gamma}_l, \tag{4.15}$$

where α'_l is given by (4.11).

We will show that (4.15) is an approximation for (4.14) by showing that

$$\frac{n}{n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log S'_k \approx \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log S_k \tag{4.16}$$

and

$$\frac{n}{n'}\sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} \alpha'_l \hat{\gamma}_l \approx \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \alpha_l \hat{\gamma}_l. \tag{4.17}$$

For (4.16) we consider the relationship between S'_k and S_k. It is easier to think about the situation where n divides n', but the arguments extend to any $n' < n$.

Replacing k by $(n'/n)k$ in the restricted spectrum definition (4.9), we have

$$S'_{\frac{n'}{n}k} = \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} \gamma_l e^{\frac{-2\pi i k l}{n}}.$$

Comparing this with the full spectrum (4.4), we have

$$\begin{aligned} S_k - S'_{\frac{n'}{n}k} &= \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \gamma_l e^{\frac{-2\pi i k l}{n}} - \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} \gamma_l e^{\frac{-2\pi i k l}{n}} \\ &= \sum_{l=-\frac{n}{2}}^{-\frac{n'}{2}-1} \gamma_l e^{\frac{-2\pi i k l}{n}} + \sum_{l=\frac{n'}{2}}^{\frac{n}{2}-1} \gamma_l e^{\frac{-2\pi i k l}{n}} \\ &\approx 0 \text{ if } \gamma_l \approx 0 \text{ for } |l| \geq \frac{n'}{2}. \end{aligned}$$

Therefore if we take n' sufficiently large, $S_k \approx S'_{\frac{n'}{n}k}$, or equivalently,

$$S_{\frac{n}{n'}k} \approx S'_k. \tag{4.18}$$

For (4.16) we can then write

$$\frac{n}{n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1}\log S'_k \approx \frac{n}{n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1}\log S_{\frac{n}{n'}k} \approx \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\log S_k,$$

where the goodness of the second approximation can be seen by considering the composite trapezoidal rule for integration (4.1). Since S_k is an even function in k we can write

$$\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\log S_k = \log S_0 + \log S_{\frac{n}{2}} + 2\sum_{k=1}^{\frac{n}{2}-1}\log S_k \tag{4.19}$$

and by considering this as a sum based on n subintervals and considering the sum based on n' subintervals, we get the approximation

$$\begin{aligned}\frac{n}{n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1}\log S'_k &= \frac{n}{n'}\left[\log S'_0 + \log S'_{\frac{n'}{2}} + 2\sum_{k=1}^{\frac{n'}{2}-1} S'_k\right] \\ &\approx \frac{n}{n'}\left[\log S_0 + \log S_{\frac{n}{2}} + 2\left(S_{\frac{n}{n'}} + S_{2\frac{n}{n'}} + \ldots + S_{\frac{n}{2}-\frac{n}{n'}}\right)\right],\end{aligned}$$

using (4.18). By consideration of (4.2), it can be seen that the error in this approximation is of order $(1/n')^2$. We are approximating the same sum as (4.19) but by fewer terms, so as long as S_k is a continuous function of k, (4.16) holds.

For (4.17) we use the property that the coefficients α_l decay exponentially as l increases. In particular we assume $\alpha_l \approx 0$ for $|l| > \frac{n'}{2}$. Then we simply need to show that

$$\frac{n}{n'}\alpha'_l \approx \alpha_l.$$

The terms being summed in (4.17) are then the same each side, the left hand side just omitting some terms for which $\alpha_l \approx 0$.

Putting (4.18) into (4.11) we have

$$\frac{n}{n'}\alpha'_l = \frac{n}{n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1}\frac{1}{S'_k}e^{\frac{2\pi ikl}{n'}} \approx \frac{n}{n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1}\frac{1}{S_{\frac{n}{n'}k}}e^{\frac{2\pi ikl}{n'}} \approx \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\frac{1}{S_k}e^{\frac{2\pi ikl}{n}} = \alpha_l.$$

The second approximation again utilises the compound trapezoidal rule for integration and depends on the continuity of the function being summed. Now (4.17) holds and we have shown that under the conditions assumed, the restricted likelihood of (4.8) provides an approximation to the full likelihood of (4.7).

4.3 Multivariate series

4.3.1 Notation

The log-likelihood $\mathcal{L}$ for a multivariate stationary Gaussian mean-corrected time series of length n and dimension R, i.e. $y_r = (y_{r0}, y_{r1}, ..., y_{r,n-1})$, $r = 1, \ldots, R$, has the form of a set of independent Wishart distributions (Brillinger, 1975, page 238) and hence can be written

$$\mathcal{L} = -\frac{1}{2}\sum_{k=0}^{n-1} \log|S_k| - \frac{1}{2}\sum_{k=0}^{n-1} \text{trace}\,[S_k^{-1}\hat{S}_k], \tag{4.20}$$

where S_k and $\hat{S}_k$ are respectively the $R \times R$ complex matrices of cross-spectral and cross-periodogram coefficients at frequency $2\pi k/n$. $|\cdot|$ denotes a determinant.

Define the cross-spectrum $(S_k)_{rs}$ and cross-periodogram $(\hat{S}_k)_{rs}$ at frequency $2\pi k/n$, between series r and s, as the discrete Fourier transform of the theoretical and sample cross-covariances, respectively, i.e.

$$(S_k)_{rs} = \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} (\gamma_l)_{rs} e^{\frac{-2\pi i k l}{n}} \quad \text{for } k = 0, 1, \ldots, n-1, \tag{4.21}$$

$$(\hat{S}_k)_{rs} = \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} (\hat{\gamma}_l)_{rs} e^{\frac{-2\pi i k l}{n}} \quad \text{for } k = 0, 1, \ldots, n-1, \tag{4.22}$$

where $i = \sqrt{-1}$ and the sample cross-covariance at lag l, between series r and s, is defined, assuming a circular model, as

$$(\hat{\gamma}_l)_{rs} = \frac{1}{n}\sum_{j=0}^{n-1} y_{rj} y_{s,j+l \bmod n}.$$

We again assume throughout that n is even. If n is odd, the limits on the sums in (4.21) and (4.22) become $\pm(n-1)/2$ and subsequent expressions change accordingly.

As in the univariate case we have circular sums and so (4.20) can equivalently be written as

$$\mathcal{L} = -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log|S_k| - \frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \text{trace}\,[S_k^{-1}\hat{S}_k], \tag{4.23}$$

in which case we would consider (4.21) and (4.22) defined for $k = -n/2, \ldots, n/2-1$.

The restricted log-likelihood is then given by

$$\mathcal{L}' = -\frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log|S'_k| - \frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \text{trace}\,[S_k'^{-1}\hat{S}'_k], \tag{4.24}$$

where $n' < n$. S'_k and $\hat{S}'_k$ are the restricted cross-spectrum and cross-periodogram, respectively, formed from the expected and sample cross-covariances up to lag $n'/2$ only. Both are $R \times R$ matrices with (r, s)th elements, respectively

$$(S'_k)_{rs} = \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} (\gamma_l)_{rs} e^{\frac{-2\pi ikl}{n'}} \quad \text{for} \quad k = -\frac{n'}{2}, \ldots, \frac{n'}{2} - 1, \tag{4.25}$$

$$(\hat{S}'_k)_{rs} = \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1} (\hat{\gamma}_l)_{rs} e^{\frac{-2\pi ikl}{n'}} \quad \text{for} \quad k = -\frac{n'}{2}, \ldots, \frac{n'}{2} - 1.$$

We introduce a generalisation of the *inverse cross-covariance coefficient* (4.10), defined as the inverse Fourier transform of the inverse of the cross-spectrum. For each lag l this is an $R \times R$ matrix, the (r, s)th element being given by

$$(\alpha_l)_{rs} = \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} (S_k^{-1})_{rs} e^{\frac{2\pi ikl}{n}}, \tag{4.26}$$

the corresponding restricted form being

$$(\alpha'_l)_{rs} = \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} (S_k'^{-1})_{rs} e^{\frac{2\pi ikl}{n'}}, \tag{4.27}$$

i.e. the inverse Fourier transform of the inverse of the restricted cross-spectrum.

4.3.2 Full likelihood

We provide a derivation of the full likelihood in the multivariate case (4.20). It is sufficient to prove that the $\hat{S}_k$, $k = 0, \ldots, n-1$ are independent and follow a complex Wishart distribution, see, for example, Johnson and Kotz (1972, Chapter 38). The proof is along similar lines to that for the univariate case in Section 4.2.2.

Consider the quantities a_{rk} and b_{rk}, respectively the cosine and sine Fourier transforms of the univariate series y_r,

$$\begin{aligned} a_{rk} &= \sum_{j=0}^{n-1} y_{rj} \cos \frac{2\pi jk}{n}, \\ b_{rk} &= \sum_{j=0}^{n-1} y_{rj} \sin \frac{2\pi jk}{n}. \end{aligned}$$

These are linear combinations of normally distributed variables with zero mean and so are themselves normally distributed with zero mean. Using the standard

trigonometric addition formulae we can show that

$$\begin{aligned}
\mathrm{Cov}(a_{rk}, a_{sm}) &= E\left(\sum_{j=0}^{n-1} y_{rj} \cos\frac{2\pi jk}{n} \sum_{l=0}^{n-1} y_{sl} \cos\frac{2\pi lm}{n}\right) \\
&= \sum_j \sum_l \cos\frac{2\pi jk}{n} \cos\frac{2\pi lm}{n} E(y_{rj} y_{sl}) \\
&= \frac{n}{2}\delta_{km} \sum_l (\gamma_l)_{rs} \cos\frac{2\pi kl}{n} \\
&= \frac{n}{2}\delta_{km}\ \mathrm{Re}(S_k)_{rs}.
\end{aligned}$$

It can be shown in a similar way that

$$\begin{aligned}
\mathrm{Cov}(b_{rk}, b_{sm}) &= \frac{n}{2}\delta_{km} \sum_l (\gamma_l)_{rs} \cos\frac{2\pi kl}{n} = \frac{n}{2}\delta_{km}\ \mathrm{Re}(S_k)_{rs}, \\
\mathrm{Cov}(a_{rk}, b_{sm}) &= -\frac{n}{2}\delta_{km} \sum_l (\gamma_l)_{rs} \sin\frac{2\pi kl}{n} = \frac{n}{2}\delta_{km}\ \mathrm{Im}(S_k)_{rs},
\end{aligned}$$

where $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary parts of a complex number, respectively. From this we can say that for series r and s and for frequencies $2\pi k/n$ and $2\pi m/n$, the set $\{a_{rk}, b_{rk}, a_{sk}, b_{sk}\}$ are independent of $\{a_{rm}, b_{rm}, a_{sm}, b_{sm}\}$ for $k \neq m$.

Let us now define

$$f_{rk} = a_{rk} + ib_{rk} = \sum_j y_{rj} e^{\frac{2\pi ijk}{n}}.$$

Also write

$$\begin{aligned}
a_k &= (a_{1k}, a_{2k}, \ldots, a_{Rk}), \\
b_k &= (b_{1k}, b_{2k}, \ldots, b_{Rk}), \\
f_k &= (f_{1k}, f_{2k}, \ldots, f_{Rk}).
\end{aligned}$$

Then $f_k, k = 0, \ldots, n-1$ are independent and follow multivariate complex normal distributions, i.e.

$$f_k \sim N_R^C(\mathbf{0}, V_k),$$

where $\mathbf{0}$ is of length R and V_k is an $R \times R$ complex matrix.

Instead of thinking about a complex variable of dimension R we can think about the real and imaginary parts together forming a real vector of length $2R$, i.e.

$$\begin{pmatrix} \mathrm{Re}\ f_k \\ \mathrm{Im}\ f_k \end{pmatrix} = \begin{pmatrix} a_k \\ b_k \end{pmatrix} \sim N_{2R}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \frac{1}{2}\begin{pmatrix} \mathrm{Re}\ V_k & -\mathrm{Im}\ V_k \\ \mathrm{Im}\ V_k & \mathrm{Re}\ V_k \end{pmatrix}\right),$$

where

$$(\mathrm{Re}\ V_k)_{rs} = \mathrm{Cov}(a_{rk}, a_{sk}) = \mathrm{Cov}(b_{rk}, b_{sk}) = \frac{n}{2}\ \mathrm{Re}(S_k)_{rs}$$

and

$$(\text{Im } V_k)_{rs} = \text{Cov}(a_{rk}, b_{sk}) = -\text{Cov}(b_{rk}, a_{sk}) = \frac{n}{2} \text{Im}(S_k)_{rs}.$$

Now, writing the complex conjugate of f_{rk} as $\bar{f}_{rk} = a_{rk} - ib_{rk}$, we have

$$f_{rk}\bar{f}_{sk} = \sum_l (\hat{\gamma}_l)_{rs} e^{-\frac{2\pi ikl}{n}} = (\hat{S}_k)_{rs},$$

the cross-periodogram coefficient between series r and s at frequency $2\pi k/n$. So for each $k = 0, \ldots, n-1$, we have an $R \times R$ complex matrix $\hat{S}_k$, with elements

$$\begin{aligned} (\hat{S}_k)_{rs} &= f_{rk}\bar{f}_{sk} \\ \iff \quad \hat{S}_k &= f_k \bar{f}_k{}^T. \end{aligned}$$

Then by, for example, Brillinger (1975, page 238) or Johnson and Kotz (1972, Chapter 38), we have that $\hat{S}_k$ follows a complex Wishart distribution of dimension R with a single degree of freedom. The density function for this distribution is proportional to

$$\begin{aligned} & |V_k|^{-1} \exp\{-\text{trace } (V_k^{-1}\hat{S}_k)\} \\ \propto \quad & |S_k|^{-1} \exp\{-\text{trace } (S_k^{-1}\hat{S}_k)\} \end{aligned}$$

since

$$V_k = \frac{n}{2} S_k.$$

Hence f_{rk} and f_{sm} are independent for $k \neq m$ and it follows that $\hat{S}_k$ and $\hat{S}_m$ are independent for $k \neq m$, and so the log likelihood for $\hat{S}_0, \ldots, \hat{S}_{n-1}$ is given by (4.20).

4.3.3 Restricted likelihood

We will show that for short-memory processes such as ARMA processes, $\mathcal{L}$, given by (4.23) can be approximated by $\mathcal{L}'$, given by (4.24).

We can write (4.23) more fully as

$$\begin{aligned} \mathcal{L} &= -\frac{1}{2} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log |S_k| - \frac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} [(S_k^{-1})_{rs} (\hat{S}_k)_{sr}] \\ &= -\frac{1}{2} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log |S_k| - \frac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} (S_k^{-1})_{rs} \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} (\hat{\gamma}_l)_{sr} e^{\frac{-2\pi ikl}{n}} \quad \text{using (4.22)} \\ &= -\frac{1}{2} \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log |S_k| - \frac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \left[\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} (S_k^{-1})_{rs} e^{\frac{-2\pi ikl}{n}} \right] (\hat{\gamma}_l)_{sr}. \end{aligned}$$

The inverse cross-spectrum is Hermitian, therefore the result of the Fourier transform in the square bracket is real, making it equivalent to the definition of the inverse cross-covariance coefficient (4.26). Hence we can write

$$\begin{aligned}\mathcal{L} &= -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\log|S_k| - \frac{1}{2}\sum_{r=1}^{R}\sum_{s=1}^{R}\sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1}(\alpha_l)_{rs}(\hat{\gamma}_l)_{sr} \qquad (4.28)\\ &= -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\log|S_k| - \frac{1}{2}\sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1}\text{trace }[\alpha_l\hat{\gamma}_l]. \qquad (4.29)\end{aligned}$$

Similarly for $\mathcal{L}'$, we write (4.24) as

$$\mathcal{L}' = -\frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1}\log|S'_k| - \frac{n}{2n'}\sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1}\text{trace }[\alpha'_l\hat{\gamma}_l], \qquad (4.30)$$

where the entries of α'_l are given by (4.27).

To show that (4.30) is an approximation for (4.29) we will show that

$$\frac{n}{n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1}\log|S'_k| \approx \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1}\log|S_k| \qquad (4.31)$$

and

$$\frac{n}{n'}\sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1}(\alpha'_l)_{rs}(\hat{\gamma}_l)_{sr} \approx \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1}(\alpha_l)_{rs}(\hat{\gamma}_l)_{sr}. \qquad (4.32)$$

For (4.31) we need to look at the relationship between $(S'_k)_{rs}$ and $(S_k)_{rs}$. It is easier to think about the situation where n divides n', but the arguments extend to any $n' < n$.

Replacing k by $\frac{n'}{n}k$ in (4.25) we have

$$\left(S'_{\frac{n'}{n}k}\right)_{rs} = \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1}(\gamma_l)_{rs}e^{\frac{-2\pi ikl}{n}}.$$

Looking at the difference between this and (4.21) we have

$$\begin{aligned}(S_k)_{rs} - (S'_{\frac{n'}{n}k})_{rs} &= \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1}(\gamma_l)_{rs}e^{\frac{-2\pi ikl}{n}} - \sum_{l=-\frac{n'}{2}}^{\frac{n'}{2}-1}(\gamma_l)_{rs}e^{\frac{-2\pi ikl}{n}}\\ &= \sum_{l=-\frac{n}{2}}^{-\frac{n'}{2}-1}(\gamma_l)_{rs}e^{\frac{-2\pi ikl}{n}} + \sum_{l=\frac{n'}{2}}^{\frac{n}{2}-1}(\gamma_l)_{rs}e^{\frac{-2\pi ikl}{n}}\\ &\approx 0 \text{ if } (\gamma_l)_{rs} \approx 0 \text{ for } |l| \geq \frac{n'}{2}, \qquad (4.33)\end{aligned}$$

since the elements of γ_l decay exponentially. Therefore if we take n' sufficiently large, $(S_k)_{rs} \approx (S'_{\frac{n'}{n}k})_{rs}$, or equivalently,

$$(S_{\frac{n}{n'}k})_{rs} \approx (S'_k)_{rs},$$

from which it follows that

$$|S_{\frac{n}{n'}k}| \approx |S'_k|.$$

We can then write (4.31) as

$$\frac{n}{n'} \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log|S'_k| = \frac{n}{n'} \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log|S_{\frac{n}{n'}k}|, \tag{4.34}$$

and comparing this with the right hand side of (4.31), as long as $|S_k|$ is a continuous function of k then, by the compound trapezoidal rule for integration, we can approximate the right hand side of (4.31) by the right hand side of (4.34), with error of order $(1/n')^2$.

For (4.32), we assume that α decays exponentially, and therefore for sufficiently large n', $(\alpha_l)_{rs} \approx 0$ for $|l| > \frac{n'}{2}$. So, we simply need to show that

$$\frac{n}{n'}(\alpha'_l)_{rs} \approx (\alpha_l)_{rs}.$$

From the definitions of α (4.26) and α' (4.27), and using (4.33), we have

$$\begin{aligned} \frac{n}{n'}(\alpha'_l)_{rs} = \frac{n}{n'} \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} ({S'_k}^{-1})_{rs} e^{\frac{2\pi i l k}{n'}} &\approx \frac{n}{n'} \sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} ({S_{\frac{n}{n'}k}}^{-1})_{rs} e^{\frac{2\pi i l k}{n'}} \\ &\approx \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} ({S_k}^{-1})_{rs} e^{\frac{2\pi i k l}{n}} = (\alpha_l)_{rs} \end{aligned}$$

where the second approximation again depends on the continuity of the function being summed, the error being of order $(1/n')^2$.

Therefore (4.32) holds and we have shown that the full likelihood of (4.23) can be approximated by the restricted likelihood of (4.24).

4.4 Summary

It has been shown that for a multivariate stationary Gaussian process, the likelihood can be expressed in a spectral representation that is asymptotically equivalent to the conventional form. It has also been shown that the full spectral

likelihood can be approximated by a restricted form, using periodogram coefficients based on only the first few lags of cross-correlation coefficients. The precise number of lags to be retained is only described here in terms of taking a value of n' above which both the cross-covariance and inverse cross-covariance coefficients are approximately zero. This was further investigated for particular examples of ARMA processes in Chapter 3. Proofs of results have been presented separately for univariate and multivariate series. Both follow similar lines of thought, but if only the univariate case is required then, as would be expected, the situation is somewhat simpler.

Chapter 5

Hidden Markov models

I review the properties of hidden Markov models (HMMs) and consider their suitability for modelling behaviour data, focusing on the cow feeding data. After considering the basic form of the model in Section 5.1, Section 5.2 introduces the notation used, as well as looking at the form of the likelihood, ways of incorporating diurnal pattern into the model, issues of model selection and ways of recovering the latent states after a model has been selected. In Section 5.3, HMMs are fitted to the cow feeding data and issues of parameter redundancy, model selection and model diagnostics are discussed.

5.1 Motivation

Hidden Markov models appear extensively in the electronic engineering literature, being used as models for speech processing and recognition, see for example the extensive review in Rabiner (1989). The basic model comprises a set of unobserved states following a Markov chain, plus a set of observations which are conditionally independent given the underlying states. The model can be represented by the conditional independence graph of Figure 5.1 (MacDonald and Zucchini, 1997, page 67), where C_t is the underlying state at time t and X_t is the observation at time t. This is an attractive model for behaviour data because, as previously discussed, the rapid changes in observed behaviour may not be as important as the underlying motivational state of the animal, which is forming the basis of the model here. For the feeding data it is intuitively appealing to have a feeding state which also contains the short gaps between observed feeding. These short gaps can clearly be considered as part of the overall meal, occurring only due to the cow moving between feeders or taking a short break. So although the cow is observed to change behaviour, it is likely that this is merely an artifact

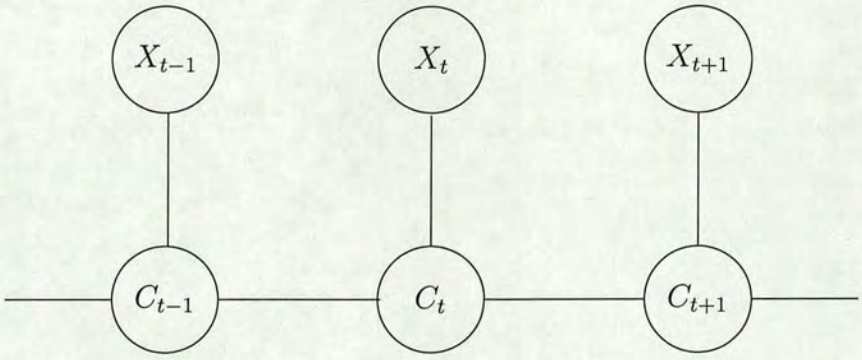

Figure 5.1: *Conditional independence graph for a hidden Markov model. C_t is the underlying state and X_t the observed behaviour at time t.*

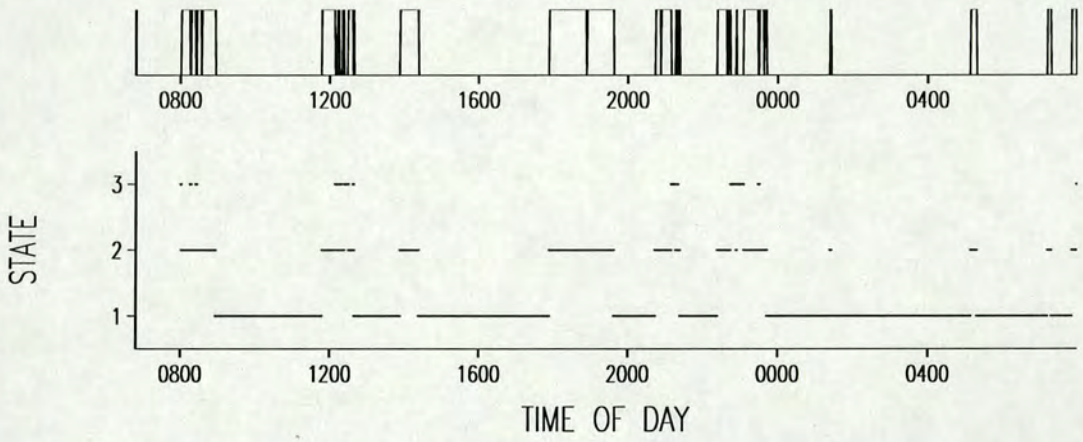

Figure 5.2: *Simulated data from a three-state HMM with parameters estimated as for Cow 108. Lower plot shows current state, either 1, 2 or 3; upper plot shows the corresponding feeding events that would be observed.*

of the mechanics of feeding rather than a change in the underlying motivational state of the animal. Figure 5.2 shows an example of simulated data for a hidden Markov model with three states. State 1 is clearly a non-feeding state; in particular it is the between-meal state. State 2 is generally feeding, but includes some of the shorter within-meal gaps, and State 3 allows for the longer within-meal gaps.

MacDonald and Raubenheimer (1995) introduced HMMs to the behavioural literature, using them to model the perambulatory behaviour of locusts. They put forward the idea that in terms of model parsimony, a hidden Markov model can be a better alternative to a higher order Markov model for data for which a first order Markov model is an inadequate description. Nevertheless it is the direct

modelling of the underlying state of the animal that is the main motivation for the consideration of HMMs here.

It should be noted that hidden Markov models exist only in a discrete-time framework. This might be considered somewhat restrictive, as our data clearly occurs naturally in continuous time (see Section 1.4.1 for more discussion of this). Nevertheless it can be conceived that as long as the size of the discretisation unit is sufficiently small, for example the same as the smallest unit of time that behaviour can be observed at, then this is not a problem. This point is discussed again in the context of all models fitted in Section 7.1.

5.2 Theory

In this section we outline some theory of HMMs. The notation adopted is similar to that used by MacDonald and Zucchini (1997, page 65).

5.2.1 Notation

Recall that X_t denotes the observed behaviour and C_t the underlying state at time t. Let $\{C_t : t = 1, \ldots, n\}$ be an irreducible homogeneous stationary Markov chain on the state space $\{1, 2, \ldots, K\}$, with *transition probability matrix* Γ with elements

$$\Gamma_{kl} = P(C_t = l | C_{t-1} = k).$$

The rows of the transition matrix must sum to zero, therefore we have the constraints

$$\sum_{l=1}^{K} \Gamma_{kl} = 1 \quad \text{for} \quad k = 1, \ldots, K.$$

The *stationary distribution* δ, where $\delta' = (\delta_1, \ldots, \delta_K)$, can be completely determined by solving the linear system $\delta'\Gamma = \delta'$, subject to the constraint $\sum_{k=1}^{K} \delta_k = 1$.

We next define the *state-dependent probabilities*, π. These are the probabilities of observing behaviour x conditional on being in state k, given by

$$\pi_{xk} = P(X_t = x | C_t = k) \quad \text{for} \quad k = 1, \ldots, K. \tag{5.1}$$

Defining a matrix

$$\lambda(x) = \text{diag}(\pi_{x1}, \pi_{x2}, \ldots, \pi_{xK}), \tag{5.2}$$

the likelihood of the observed series $x_1, x_2, \ldots, x_n$ can be written as

$$\mathcal{L} = \delta' \, \Gamma \, \lambda(x_1) \, \Gamma \, \lambda(x_2) \ldots \Gamma \, \lambda(x_n) \, \mathbf{1}, \tag{5.3}$$

where $\mathbf{1}$ is a column vector of length K with each element equal to 1. The likelihood can be seen to be a simple product of matrices — the stationary distribution for the first state, followed by a product of transition probabilities to subsequent states and state-dependent probabilities for the observation given each state. The ability to write down the likelihood explicitly is one of the appealing properties of HMMs. For models with a more general conditional independence graph than the one in Figure 5.1, i.e. with a more complicated dependence structure, for example the current observation dependent on both the current and previous states, an explicit form for the likelihood cannot in general be written down.

5.2.2 Evaluation and maximisation of the likelihood

In theory the likelihood can be maximised directly, but in practice there are problems due to the number of calculations involved and with underflow, since (5.3) is a product of matrices and therefore additive, forcing us to work with the likelihood itself and not the log-likelihood. We therefore consider the *forward-backward algorithm* as a way of evaluating the likelihood. This is described, for example, in Lindgren (1978), also in Le et al. (1992) as a reaction to the methodology used by Albert (1991), who incorrectly claimed that computations are too intensive to allow the likelihood to be evaluated exactly by any method and hence uses an approximation.

The forward-backward algorithm is so-called because it entails two passes through the data, one in each direction. In detail, we have the *forward probabilities*, α_{tk}, and the *backward probabilities*, β_{tk}, given by

$$\begin{aligned} \alpha_{tk} &= P(X_1 = x_1, \ldots, X_t = x_t, C_t = k) \\ \beta_{tk} &= P(X_{t+1} = x_{t+1}, \ldots, X_n = x_n | C_t = k) \end{aligned}$$

for times $t = 1, \ldots, n$ and states $k = 1, \ldots, K$. So α_{tk} is the joint probability of the observed series up to time t and the state at time t being k; β_{tk} is the conditional probability of the observed series from time $t+1$ onwards, given that the state at time t is k. These are calculated recursively for each state k separately, the α's in the forward direction, $t = 1, \ldots, n$, and the β's in a backward direction, $t = n, n-1, \ldots, 1$.

$$\alpha_{1k} = P(C_1 = k) P(X_1 = x_1 | C_1 = k)$$

$$
\begin{aligned}
&= \delta_k P(X_1 = x_1 | C_1 = k), \\
\alpha_{tk} &= \sum_{l=1}^{K} \alpha_{t-1,l} \Gamma_{lk} P(X_t = x_t | C_t = k) \quad \text{for} \quad t = 2, \ldots, n, \\
\beta_{nk} &= 1, \\
\beta_{tk} &= \sum_{l=1}^{K} \beta_{t+1,l} \Gamma_{kl} P(X_{t+1} = x_{t+1} | C_{t+1} = l) \quad \text{for} \quad t = n-1, n-2, \ldots, 1.
\end{aligned}
$$

The log-likelihood is then given by

$$\mathcal{L} = \log \sum_{k=1}^{K} \alpha_{tk} \beta_{tk}$$

for any t. In particular, for $t = n$,

$$\mathcal{L} = \log \sum_{k=1}^{K} \alpha_{nk}. \tag{5.4}$$

Hence if we purely want to evaluate the likelihood at given parameter values, calculation of the β's is not required.

Scaling of the calculations is a problem throughout, as the recursive probabilities that need to be calculated soon become too small for a computer to hold. The suggestion of MacDonald and Zucchini (1997, page 79) is to scale the α's so that at each stage, their average value is 1. The overall scaling factor can be stored cumulatively on the log-scale and then added to (5.4) at the end. Maximisation of the likelihood can be carried out numerically by a routine such as E04JAF (Numerical Algorithms Group, 1993), a quasi-Newton algorithm for minimising a function of several variables using function values only and allowing simple bounds on the parameters. For a model with only two states this is straightforward. For more than two states there are a further K constraints and so we must either use a more general routine such as E04UCF which allows non-linear constraints on the parameters, or else reparameterise so that simple bounds and hence E04JAF can be used.

In the speech processing literature, rather than maximise the likelihood directly, the *Baum-Welch algorithm* is generally used to estimate parameters. This algorithm, developed over a number of papers, e.g. Baum and Eagon (1967) and Baum et al. (1970), is a form of the EM algorithm (see also Section 6.3) for which the complete data consist of both the states and the observations, and the missing data are the states. The forward-backward algorithm still needs to be used within the expectation step to evaluate all the necessary conditional probabilities, and then the Baum-Welch algorithm re-estimates the parameters — essentially we have three re-estimation equations, for δ, π and Γ respectively. It can be

shown that this algorithm guarantees an improvement in the likelihood, except at critical points. Note that here δ is estimated separately from Γ; stationarity is not assumed and δ is considered to be the initial distribution of the states, not the stationary distribution. Here we assume stationarity and take the approach of maximising the likelihood numerically rather than using the Baum-Welch algorithm. In the case $K = 2$ the approaches are identical.

5.2.3 Diurnal pattern

We have seen from the data in Appendix A and in Section 2.4 that for some cows there is a clear diurnal pattern. This can be allowed for in a hidden Markov model in two ways, either via the transition probabilities or via the state-dependent probabilities. Dealing with the latter of these first, we modify the probability of behaviour X occurring in state k, π_{xk}, to become $\pi_{xk}(t)$, so instead of estimating this overall, we now model it in terms of time t. The natural way to model a probability is via the logit transformation, for example,

$$\text{logit } \pi_{xk}(t) = \log[\pi_{xk}(t)/(1 - \pi_{xk}(t))] = f(t).$$

This allows for an overall trend with time via the function $f(t)$, which typically will be some parametric function of t, e.g. linear or sinusoidal. In Section 2.4 we saw that it was difficult to allow for the trend adequately with a parametric form, and so we allowed a categorical time trend on an hourly basis. We can do similarly here and allow $f(t)$ to take a different value for each hour of the day. In any case, the form of the likelihood for the HMM does not change — the constant π_{xk} above is simply replaced by $\pi_{xk}(t)$, given by the inverse transformation:

$$\pi_{xk}(t) = \exp(f(t))/[1 + \exp(f(t))].$$

Parameters for the time trend can either be estimated separately for each state or can be constrained to be the same for all states.

To incorporate time trend via the transition probabilities, we model them in an analogous way to the state-dependent probabilities above. Now the Markov chain is no longer homogeneous and there is no overall stationary distribution. Instead we assume an initial distribution at time $t = 1$, which must be estimated in addition to the transition probabilities and state-dependent probabilities. The estimation can be done by maximisation of the likelihood conditional on the initial state of the Markov chain. MacDonald and Zucchini (1997, pages 130–133) give more details.

5.2.4 Model selection

The main issue here is that of the number of underlying states. The approach is to fit models with increasing numbers of states and use some criterion to decide whether the additional states are needed. This might seem a situation in which to use the generalised likelihood ratio test, relating the difference between maximised log-likelihood values to chi-squared distributions, but the regularity conditions for the required asymptotic theory do not hold here. The situation is analogous to the problem of determining the number of components in a mixture and the asymptotic theory breaks down because the null hypothesis being tested is whether one of the mixing parameters is zero, which is on the boundary of the parameter space (Titterington, 1990). Widely used alternative approaches are to use either Akaike's information criterion, AIC (Akaike, 1974) or Schwarz's Bayesian information criterion, BIC (Schwarz, 1978), given respectively by

$$\begin{aligned} AIC &= -2\mathcal{L} + 2m_K, \\ BIC &= -2\mathcal{L} + m_K \log n, \end{aligned}$$

where $\mathcal{L}$ is the maximised log-likelihood for the n observations and m_K is the number of parameters estimated for the K-state model. This forms a penalised likelihood test, the model with the lower value of the criterion being the favoured one according to that criterion. Akaike's information criterion is based on the fact that because the same data are being used both to estimate parameters and calculate the likelihood, $\mathcal{L}/n$ is a biased estimator for the log-likelihood of the data given the maximum likelihood parameters; the expectation of the bias is $-m_K/n$, leading to the formula given for AIC. AIC can sometimes overestimate the number of parameters needed in a model, and from this point of view, BIC is to be preferred. This can be viewed as an asymptotic approximation to the use of Bayes factors. More about this is discussed in Section 7.2.2, and a full discussion can be found in Kass and Raftery (1995).

It should be borne in mind that as with the likelihood ratio test, strictly these criteria are invalid for comparing hidden Markov models, again because of the lack of validity of the asymptotic theory. Nevertheless use of them here will be seen to still be useful. A more thorough approach would be to use bootstrapping, as McLachlan (1987) does for components in a simple mixture model. However for a simple appraisal of the models we consider the information criteria described above and inspect graphs and marginal distributions, etc. It should be noted that of course when comparing models using the two criteria defined above, the same model is not always selected by both criteria. When this is the case, as long

as the number of observations is greater than seven, BIC results in the more parsimonious model.

5.2.5 Recovery of states

Once a hidden Markov model has been fit to data, it will often be of interest to recover the underlying states that are being assumed to have produced the observed data. This can be done in two ways, for example see Rabiner (1989). Firstly, the *Viterbi algorithm* considers the joint probability of the whole series to decide on the most likely sequence of states. This requires a recursive technique similar to the calculation of the forward probabilities α_{tk} above. Secondly, the pointwise probability can be calculated for each point of the series in turn, giving probabilities of each state at each timepoint, conditional on the rest of the observed series. Calculation of these probabilities is straightforward, given by

$$\begin{aligned} \gamma_{tk} &= P(C_t = k | X_1, \ldots, X_n, \text{parameters}) \\ &= \frac{\alpha_{tk}\beta_{tk}}{\sum_{k=1}^{K} \alpha_{tk}\beta_{tk}} \end{aligned} \tag{5.5}$$

for times $t = 1, \ldots, n$ and states $k = 1, \ldots, K$. By taking the state at time t as that corresponding to the maximum of the γ_{tk} over the K states for each timepoint, we can come up with a most likely sequence of states for the whole series. However this might result in a sequence of states which is not even valid according to the transition probabilities, and so if this is what is required the Viterbi algorithm is preferable. For illustration we use pointwise probability plots as calculated by (5.5).

5.3 Fitting to data

As already discussed, hidden Markov models exist only in a discrete time framework and so for fitting to the cow feeding data we must first discretise the timescale. We do this on a minutely basis to be consistent with what was done for the fitting of the latent Gaussian models of Chapter 3.

The state-dependent probabilities of equation (5.1) here are simply Bernoulli variables, i.e.

$$\pi_{xk} = {p_k}^x (1 - p_k)^{1-x}$$

Cow	$\hat{\Gamma}$	$\hat{\delta}'$	$\hat{p}'$
5	$\begin{pmatrix} .9925 & .0075 \\ .0657 & .9343 \end{pmatrix}$	(.8974, .1026)	(.0000, .9452)
41	$\begin{pmatrix} .9905 & .0095 \\ .0877 & .9123 \end{pmatrix}$	(.9021, .0979)	(.0000, .9463)
108	$\begin{pmatrix} .9904 & .0096 \\ .0678 & .9322 \end{pmatrix}$	(.8757, .1243)	(.0000, .9604)
169	$\begin{pmatrix} .9935 & .0065 \\ .0756 & .9244 \end{pmatrix}$	(.9208, .0792)	(.0000, .9504)
170	$\begin{pmatrix} .9913 & .0087 \\ .0737 & .9263 \end{pmatrix}$	(.8944, .1056)	(.0000, .9108)
182	$\begin{pmatrix} .9915 & .0085 \\ .0742 & .9258 \end{pmatrix}$	(.8975, .1025)	(.0000, .9521)
194	$\begin{pmatrix} .9914 & .0086 \\ .0765 & .9235 \end{pmatrix}$	(.8993, .1007)	(.0000, .9428)
221	$\begin{pmatrix} .9889 & .0111 \\ .0559 & .9441 \end{pmatrix}$	(.8338, .1662)	(.0000, .9200)

Table 5.1: *Parameter estimates for two-state HMMs.* $\hat{\Gamma}$ *are the transition probabilities,* $\hat{\delta}$ *the overall stationary distributions and* $\hat{p}$ *the state-dependent probabilities of feeding.*

for $x = 0, 1$, where p_k is the probability of feeding in state k. The models were fit as described in Section 5.2.2, using the forward-backward algorithm for evaluation of the likelihood, with scaling factors as described, and by direct numerical maximisation of the likelihood using NAG optimisation routine E04JAF (Numerical Algorithms Group, 1993).

5.3.1 Two-state models

The simplest HMM we fit has two states, in both of which feeding and non-feeding are possible. We need to estimate two transition probabilities, giving the 2×2 transition matrix Γ, and the two state-dependent probabilities of feeding, $p' = (p_1, p_2)$, giving a total of four parameters. The entries of the transition matrix determine the overall stationary probabilities, δ. Table 5.1 shows all estimates for the eight high-protein cows.

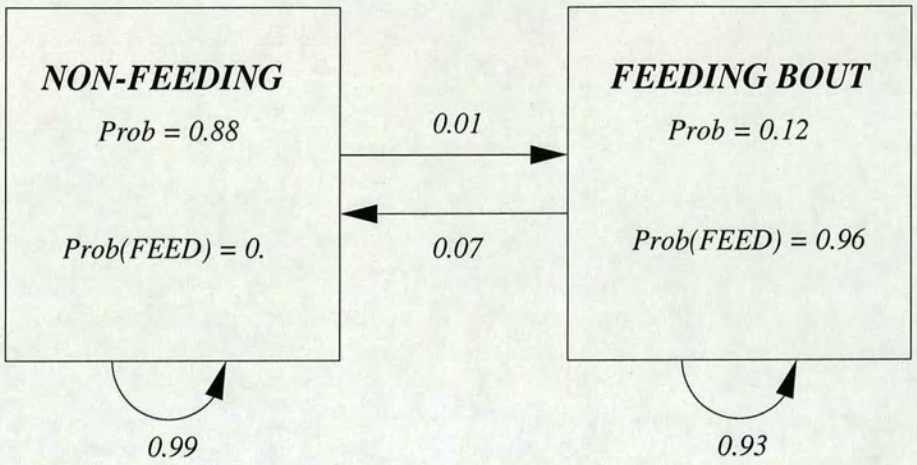

Figure 5.3: *Pictorial representation of the two-state HMM, with parameters as estimated for Cow 108.*

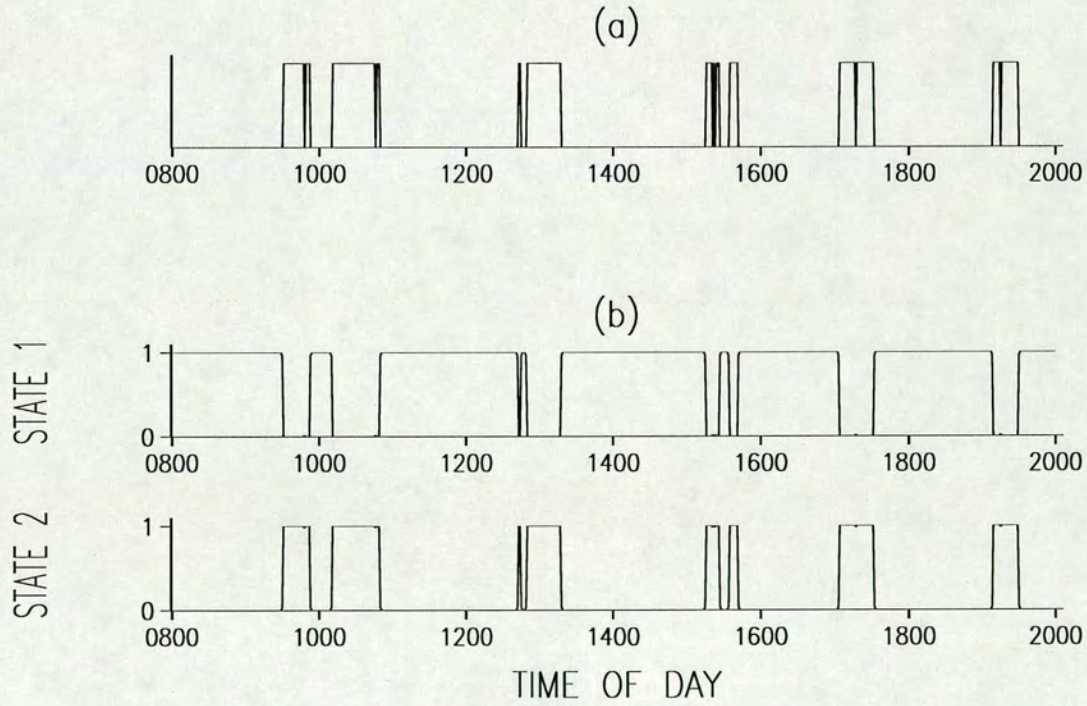

Figure 5.4: *Cow 108; (a) sample of data, (b) pointwise probabilities of being in each of the two states for the two-state HMM shown in Figure 5.3.*

The two-state HMM has a particularly attractive interpretation here biologically. The first state is always a state for which there is zero probability of feeding, hence this is the state a cow is in between feeding bouts. The other state is nominally a feeding state, with a probability of feeding generally around 0.95, so this single state can be thought of as corresponding to feeding bouts and, as such, allows for both the feeding events themselves and the short gaps separating them. Figure 5.3 is a pictorial representation of the model, with parameters as estimated for Cow 108. Figure 5.4 shows a sample of data from Cow 108 along with the pointwise probabilities for being in each of the two states, calculated from (5.5). The first state has zero probability of feeding and so, as expected, completely describes the between-meal periods, also including some of the longer within-meal non-feeding periods. The second state, with its probability of feeding of 0.96, completely describes the feeding events, and also covers the shorter within-meal non-feeding periods. Ideally we would have liked this second state to fully describe the feeding bouts, but it turns out that some of the longer gaps within meals are best described by the other state.

5.3.1.1 Diurnal pattern

We can allow for the diurnal pattern of feeding by letting the state-dependent probabilities of feeding vary with time. We have already seen in Section 2.4 that a simple parametric form is inadequate, so here we use a discrete approach, allowing the probability of feeding to take a different value for each hour. Doing this independently for each state would mean having 48 parameters for the state-dependent probabilities, i.e.

$$\text{logit } p_k(t) = \begin{cases} a_k & \text{for } t = 1 \\ a_k + b_{k,t} & t = 2, \ldots, 24, \end{cases}$$

for each state $k = 1, 2$.

In order to reduce this number we try two different approaches. Firstly we set the time-trend to be equal for the two states, so using only 25 parameters instead of 48, hence

$$\text{logit } p_k(t) = \begin{cases} a_k & \text{for } t = 1 \\ a_k + b_t & t = 2, \ldots, 24, \end{cases} \tag{5.6}$$

for each state $k = 1, 2$.

Secondly, having seen that the probability of feeding in state 1, the 'non-feeding' state, is always low, we keep this constant over the whole day and let the probability of feeding in the other state vary freely. This again uses 25 parameters

for the state-dependent probabilities; for state 1, we have logit $p_1(t) = a_1$, for all t, and for state 2, the probabilities are given by (5.6). Use of either of these methods to allow for time trend increases the total number of parameters to be estimated from 4 to 27.

Table 5.2 shows results for Cow 5, comparing the models fit allowing for time trend to that fit previously which ignores it. We see that the two time-dependent models only increase the log-likelihood from -3544 to -3534 and -3532 respectively, fairly small increases. The lowest value of both AIC and BIC, and hence the best model by either criterion, is the one ignoring trend. This is a combination of the likelihood increasing only slightly after allowing for time-trend, and the large increase in the number of parameters to be estimated. The situation is similar for the other cows, e.g. Cow 108, which displays more of a diurnal pattern than Cow 5. Here use of the second model only achieves an increase in log-likelihood from -4110.00 to -4082.65. Again, by both AIC and BIC, this increase is not enough to warrant the estimation of the extra parameters.

As already discussed, an alternative way to incorporate time trend in a HMM is via the transition probabilities. These are modified in a similar way, but now we need to modify two probabilities for each hour of the day, resulting in an extra 46 parameters. This approach is not pursued here, as it is unlikely to result in a large enough increase in the likelihood to outweigh the large number of parameters. However if a suitable parametric form for the trend could be established, modification of the transition probabilities is perhaps a more biologically plausible way to allow for diurnal pattern than the modification of the state-dependent probabilities investigated above. In our case however the lack of an obvious parametric form and the relatively small increase in likelihood with even the large numbers of parameters considered means that we will not consider the explicit modelling of the diurnal effect any further here.

5.3.2 Three-state models

We now consider the extension of the model to three states, and use AIC and BIC to see whether there is justification for a third state. With three states there are now six transition probabilities to estimate, and three state-dependent probabilities of feeding, a total of nine parameters.

Table 5.3 shows parameter estimates for the three-state models. It is noticeable that some of the transition probabilities are estimated as being very close to zero. Therefore we suspect that the model may be over-parameterised, in which case

	No trend	*Same trend in both states*	*Trend in one state only*
$\hat{\Gamma}$	$\begin{pmatrix} .9925 & .0075 \\ .0657 & .9343 \end{pmatrix}$	$\begin{pmatrix} .9925 & .0075 \\ .0654 & .9346 \end{pmatrix}$	$\begin{pmatrix} .9926 & .0074 \\ .0653 & .9347 \end{pmatrix}$
$\hat{\delta}'$	(.8974, .1026)	(.8977, .1023)	(.8977, .1023)
$\hat{p}'$	(.0000, .9452)		(.0001,)
$t = 1$		(.0001, .9460)	(.9260)
2		(.0001, .9495)	(.9619)
3		(.0001, .9629)	(.9651)
4		(.0001, .9389)	(.9331)
5		(.0002, .9701)	(.9703)
6		(.0001, .9630)	(.9586)
7		(.0002, .9723)	(.9667)
8		(.0001, .9208)	(.9143)
9		(.0001, .9450)	(.9438)
10		(.0001, .9485)	(.9490)
11		(.0001, .9426)	(.9381)
12		(.0001, .9120)	(.8942)
13		(.0001, .9553)	(.9565)
14		(.0001, .9570)	(.9567)
15		(.0001, .9589)	(.9618)
16		(.0000, .8939)	(.8908)
17		(.0001, .9462)	(.9466)
18		(.0001, .9272)	(.9223)
19		(.0001, .9228)	(.9227)
20		(.0001, .9394)	(.9361)
21		(.0001, .9602)	(.9823)
22		(.0001, .9309)	(.9164)
23		(.0001, .9505)	(.9761)
24		(.0001, .9450)	(.9388)
$-\mathcal{L}$	3543.75	3533.85	3531.98
m	4	27	27
AIC	7095.5	7121.7	7118.0
BIC	7130.2	7355.9	7352.1

Table 5.2: *Parameter estimates and likelihood for models fit to Cow 5 allowing for time trend.* $\hat{\Gamma}$ *are the transition probabilities,* $\hat{\delta}$ *the overall stationary distributions and* $\hat{p}$ *the state-dependent probabilities of feeding, dependent on hour of day where applicable.* t *is the hour of the day, labelled from 1 to 24,* $\mathcal{L}$ *is the maximised log-likelihood and* m *is the number of parameters estimated.*

Cow	$\hat{\Gamma}$	$\hat{\delta}'$	$\hat{p}'$
5	$\begin{pmatrix} .9945 & .0000 & .0055 \\ .0498 & .8966 & .0535 \\ .0000 & .3876 & .6124 \end{pmatrix}$	(.8776, .0967, .0258)	(.0000, 1.0000, .0093)
41	$\begin{pmatrix} .9947 & .0053 & .0000 \\ .0497 & .8830 & .0673 \\ .0000 & .2327 & .7673 \end{pmatrix}$	(.8794, .0935, .0270)	(.0000, .9893, .0036)
108	$\begin{pmatrix} .9936 & .0001 & .0063 \\ .0437 & .9200 & .0363 \\ .0000 & .2062 & .7938 \end{pmatrix}$	(.8309, .1221, .0470)	(.0000, .9771, .0000)
169	$\begin{pmatrix} .9950 & .0015 & .0035 \\ .0599 & .8940 & .0460 \\ .0000 & .3691 & .6309 \end{pmatrix}$	(.9066, .0754, .0181)	(.0000, .9950, .0146)
170	$\begin{pmatrix} .9946 & .0054 & .0000 \\ .0480 & .8872 & .0648 \\ .0000 & .2809 & .7191 \end{pmatrix}$	(.8781, .0991, .0229)	(.0000, .9704, .0000)
182	$\begin{pmatrix} .9958 & .0042 & .0000 \\ .0193 & .9025 & .0782 \\ .0427 & .1514 & .8059 \end{pmatrix}$	(.8613, .0989, .0398)	(.0000, .9864, .0000)
194	$\begin{pmatrix} .9935 & .0034 & .0031 \\ .0597 & .8971 & .0433 \\ .0000 & .3374 & .6626 \end{pmatrix}$	(.8825, .0969, .0206)	(.0000, .9803, .0000)
221	$\begin{pmatrix} .9929 & .0000 & .0071 \\ .0358 & .9162 & .0481 \\ .0000 & .2827 & .7173 \end{pmatrix}$	(.7947, .1584, .0470)	(.0000, .9631, .0065)

Table 5.3: *Parameter estimates for three-state HMMs.* $\hat{\Gamma}$ *are the transition probabilities,* $\hat{\delta}$ *the overall stationary distributions and* $\hat{p}$ *the state-dependent probabilities of feeding.*

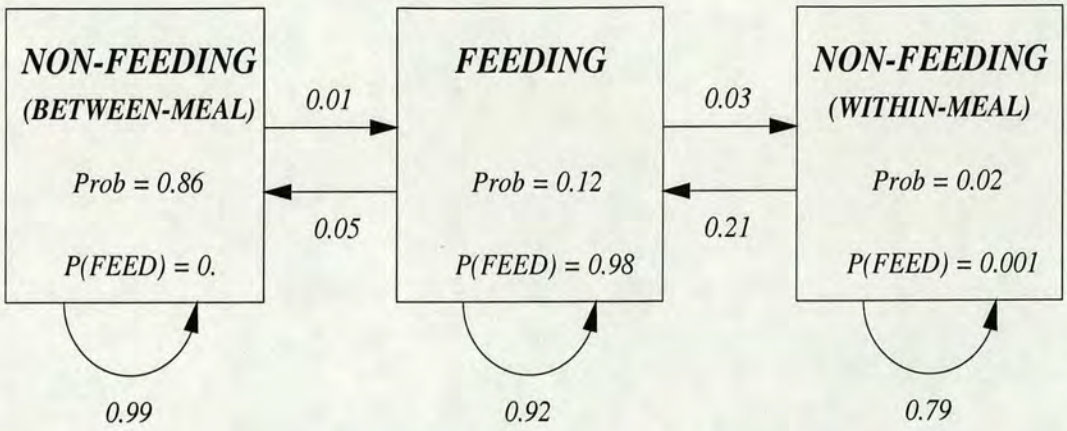

Figure 5.5: *Pictorial representation of the three-state HMM, with parameters as estimated for Cow 108.*

we would like to fix some of the parameters. The way we choose to do this, which seems sensible biologically, is to set the transition probabilities between the two states that are mainly non-feeding, states 1 and 3, to zero. Table 5.4 shows the resulting parameter estimates. Now there is a total of only seven parameters to estimate, four transition probabilities and three state-dependent probabilities of feeding.

Figure 5.5 shows this model pictorially for Cow 108, and Figure 5.6 shows the pointwise probabilities for being in each of the three states. States 1 and 2 have similar interpretations as they did for the two state model. The third state takes out some of the longer within-meal intervals from state 2.

5.3.3 Comparison of fitted HMMs

Issues of model selection were discussed in Section 5.2.4. For the models fitted above, we have estimated four parameters for the two-state models and nine or seven parameters for the three-state models. Table 5.5 shows AIC and BIC for the models for all eight high-protein cows. In all cases, both AIC and BIC decrease when the third state is added, the model with no transition between the two non-feeding states resulting in the lowest values. The likelihood for the two three-state models is almost identical, but the model with seven parameters instead of nine is clearly to be preferred in the interests of model parsimony. It can therefore be concluded that this three-state HMM with no transition probabilities between the nominally non-feeding states is the preferred one overall here.

Models with four states have not been considered. Increasing the number of states this far makes the biological interpretation of the model too complicated

Cow	$\hat{\Gamma}$	$\hat{\delta}'$	$\hat{p}'$
5	$\begin{pmatrix} .9945 & .0055 & .0000 \\ .0504 & .8960 & .0536 \\ .0000 & .3901 & .6099 \end{pmatrix}$	(.8899, .0968, .0133)	(.0000, 1.0000, .0048)
41	$\begin{pmatrix} .9947 & .0053 & .0000 \\ .0497 & .8830 & .0673 \\ .0000 & .2327 & .7673 \end{pmatrix}$	(.8794, .0935, .0270)	(.0000, .9893, .0036)
108	$\begin{pmatrix} .9936 & .0064 & .0000 \\ .0451 & .9201 & .0348 \\ .0000 & .2052 & .7948 \end{pmatrix}$	(.8573, .1221, .0207)	(.0000, .9771, .0014)
169	$\begin{pmatrix} .9950 & .0050 & .0000 \\ .0603 & .8936 & .0461 \\ .0000 & .3660 & .6340 \end{pmatrix}$	(.9149, .0756, .0095)	(.0000, .9945, .0136)
170	$\begin{pmatrix} .9946 & .0054 & .0000 \\ .0480 & .8872 & .0648 \\ .0000 & .2809 & .7191 \end{pmatrix}$	(.8781, .0991, .0229)	(.0000, .9704, .0000)
182	$\begin{pmatrix} .9958 & .0042 & .0000 \\ .0369 & .9025 & .0606 \\ .0000 & .1941 & .8059 \end{pmatrix}$	(.8702, .0989, .0309)	(.0000, .9864, .0000)
194	$\begin{pmatrix} .9935 & .0065 & .0000 \\ .0602 & .8971 & .0427 \\ .0000 & .3374 & .6626 \end{pmatrix}$	(.8909, .0969, .0123)	(.0000, .9803, .0000)
221	$\begin{pmatrix} .9929 & .0071 & .0000 \\ .0366 & .9152 & .0482 \\ .0000 & .2887 & .7113 \end{pmatrix}$	(.8149, .1586, .0265)	(.0000, .9632, .0000)

Table 5.4: *Parameter estimates for three-state HMMs, with no transition allowed between the non-feeding states (states 1 and 3).* $\hat{\Gamma}$ *are the transition probabilities,* $\hat{\delta}$ *the overall stationary distributions and* $\hat{p}$ *the state-dependent probabilities of feeding.*

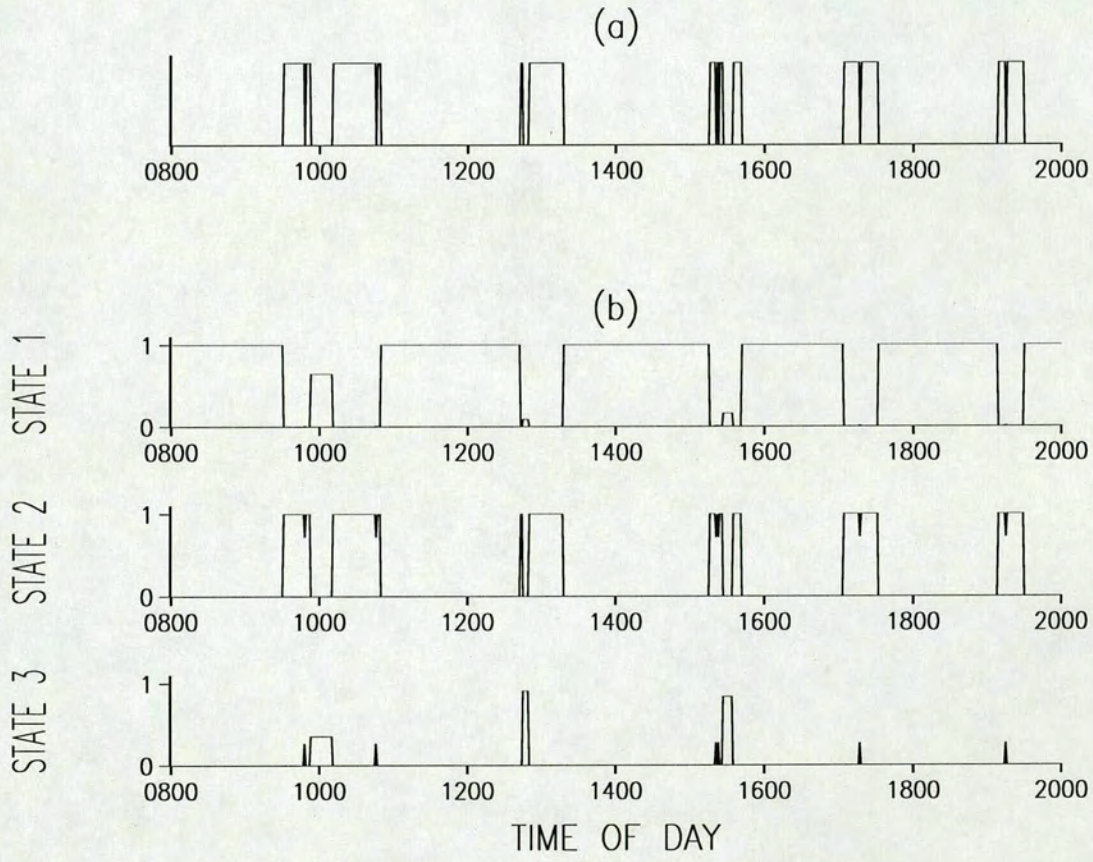

Figure 5.6: *Cow 108; (a) sample of data, (b) pointwise probabilities of being in each of the states for the three-state HMM shown in Figure 5.5.*

Cow	K	m	$-\mathcal{L}$	*AIC*	*BIC*
5	2	4	3543.75	7095.5	7130.2
	3	9	3350.57	6719.1	6797.2
	3	7	3350.91	**6715.8**	**6776.5**
41	2	4	4009.50	8027.0	8061.7
	3	9	3735.09	7488.2	7566.2
	3	7	3735.09	**7484.2**	**7544.9**
108	2	4	4110.00	8228.0	8262.7
	3	9	3973.92	7965.8	8043.9
	3	7	3973.90	**7964.8**	**8022.5**
169	2	4	3005.44	6018.9	6053.6
	3	9	2875.91	5769.8	5847.9
	3	7	2876.30	**5766.6**	**5827.3**
170	2	4	4202.18	8412.4	8447.1
	3	9	3961.91	7941.8	8019.9
	3	7	3961.91	**7937.8**	**7998.5**
182	2	4	3732.31	7472.6	7507.3
	3	9	3448.59	6915.2	6993.3
	3	7	3448.59	**6911.2**	**6971.9**
194	2	4	3838.08	7684.2	7718.9
	3	9	3713.56	7445.1	7523.3
	3	7	3713.56	**7441.1**	**7501.8**
221	2	4	5413.78	10835.6	10870.3
	3	9	5180.28	10378.6	10456.6
	3	7	5180.80	**10375.6**	**10436.3**

Table 5.5: *Log-likelihood and information criteria for two- and three-state HMMs. K is the number of states and m the number of parameters. The preferred model for each cow according to each of the criteria is highlighted.*

and far less attractive than one with fewer states. The attraction of the HMM for modelling behaviour data lies in its simplicity and the fact that it might have been possible to describe the data with just two states — meals and between meals. However one of the problems with hidden Markov models is that the states are determined by the estimation procedure and they might not turn out to have the desired practical interpretation. Even with three states we no longer have quite such a biologically elegant model, needing a separate state to allow for some of the within-meal non-feeding periods rather than being able to cope with whole meals in a single state.

5.3.4 Discrete-time compartment models

It can be noticed from the three-state hidden Markov models fitted that the models have one state that is a mainly feeding state (probability of feeding > 0.96) and the other two states are nominally non-feeding states (probability of feeding < 0.015). Therefore an obvious question to ask is how much better is the fit of the hidden Markov model than that of a model which fixes the probability of feeding in a given state at 0 or 1, as appropriate. We term this type of model a *discrete-time compartment model.* For model-fitting purposes we use the same methodology as for fitting HMMs, but now state-dependent probabilities do not need to be estimated. So for a three-state model with one feeding state and two non-feeding states, we are in effect just fitting a three-state discrete-time Markov model. In Chapter 6 we come across the corresponding continuous-time model. In either case we have the complication that for non-feeding periods, the current state is unknown — it could be either of the two non-feeding states. Therefore simple estimation procedures for discrete-time Markov models are not applicable, which is why we use the same framework as for the more general hidden Markov models above, now just having transition probabilities to estimate. A two-state model just has two transition probabilities to estimate. Three-state models have either six or four parameters to estimate, depending on whether transitions are allowed between the two non-feeding states.

Table 5.6 shows parameter estimates for the three-state model with no transition allowed between the two non-feeding states. Figure 5.7 shows this model pictorially for Cow 108, and Figure 5.8 the pointwise probability estimates for the states. It is interesting to compare the subtle differences between this latter figure and Figure 5.6, which showed the same information for the more general hidden Markov model. The main difference between these two figures are that with the discrete-time compartment model, state 2 is exclusively feeding and so

Cow	$\hat{\Gamma}$	$\hat{\delta}'$
5	$\begin{pmatrix} .9945 & .0055 & .0000 \\ .0504 & .8959 & .0537 \\ .0000 & .3927 & .6073 \end{pmatrix}$	(.8898, .0969, .0132)
41	$\begin{pmatrix} .9947 & .0053 & .0000 \\ .0508 & .8731 & .0761 \\ .0000 & .2623 & .7377 \end{pmatrix}$	(.8805, .0926, .0269)
108	$\begin{pmatrix} .9934 & .0066 & .0000 \\ .0480 & .8990 & .0531 \\ .0000 & .3185 & .6815 \end{pmatrix}$	(.8608, .1193, .0199)
169	$\begin{pmatrix} .9950 & .0050 & .0000 \\ .0607 & .8884 & .0509 \\ .0000 & .3999 & .6001 \end{pmatrix}$	(.9152, .0753, .0096)
170	$\begin{pmatrix} .9944 & .0056 & .0000 \\ .0509 & .8610 & .0881 \\ .0000 & .3584 & .6416 \end{pmatrix}$	(.8802, .0962, .0236)
182	$\begin{pmatrix} .9957 & .0043 & .0000 \\ .0382 & .8903 & .0715 \\ .0000 & .2280 & .7720 \end{pmatrix}$	(.8719, .0976, .0306)
194	$\begin{pmatrix} .9933 & .0067 & .0000 \\ .0631 & .8794 & .0575 \\ .0000 & .4446 & .5554 \end{pmatrix}$	(.8928, .0950, .0123)
221	$\begin{pmatrix} .9923 & .0077 & .0000 \\ .0415 & .8817 & .0769 \\ .0000 & .4481 & .5519 \end{pmatrix}$	(.8209, .1528, .0262)

Table 5.6: *Parameter estimates for three-state compartment models with no transition between the non-feeding states (states 1 and 3).* $\hat{\Gamma}$ *are the transition probabilities and* $\hat{\delta}$ *the overall stationary distributions.*

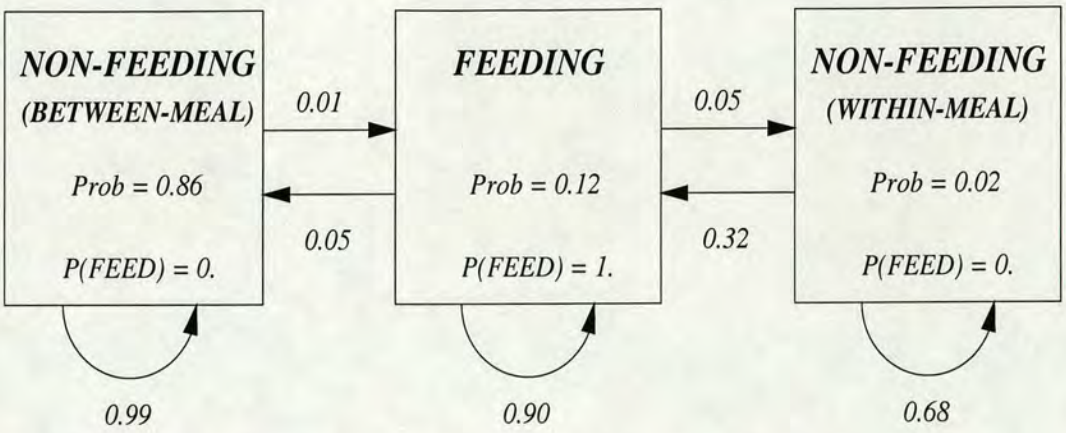

Figure 5.7: *Pictorial representation of the three-state discrete-time compartment model, with parameters as estimated for Cow 108.*

cannot include any of the within-meal intervals, hence these are now included by state 3.

We do not present parameter estimates for the more general three-state model that allows transitions between the non-feeding states, or for the two-state model, but their likelihood values are shown in Table 5.7, along with values for AIC and BIC. It can be seen from this table that, as for the HMMs, the three-state model with no transition between non-feeding states is universally the preferred model. The two-state model is too simple to describe the situation adequately, and the extra parameters in the more general three-state model produce no further increase in the likelihood from the other three-state model.

5.3.5 Comparison of hidden Markov and discrete-time compartment models

Having considered both the hidden Markov models and discrete-time compartment models separately, we now compare the two classes of model. We saw in the previous sections that universally across cows, in both cases the best model has three states and is restricted to not allow transitions between two of the states — for the hidden Markov models these are the two states that are nominally feeding, for the compartment models the states are exactly non-feeding. Table 5.8 shows values of AIC and BIC for both types of model, reproduced here from Tables 5.5 and 5.7 so that the comparisons can be made easily. It can be seen that different conclusions are drawn for different animals. For some, the compartment model gives a sufficiently good description of the data, but for others, the extra flexibility provided by the hidden Markov model is needed. Where the criteria disagree, as

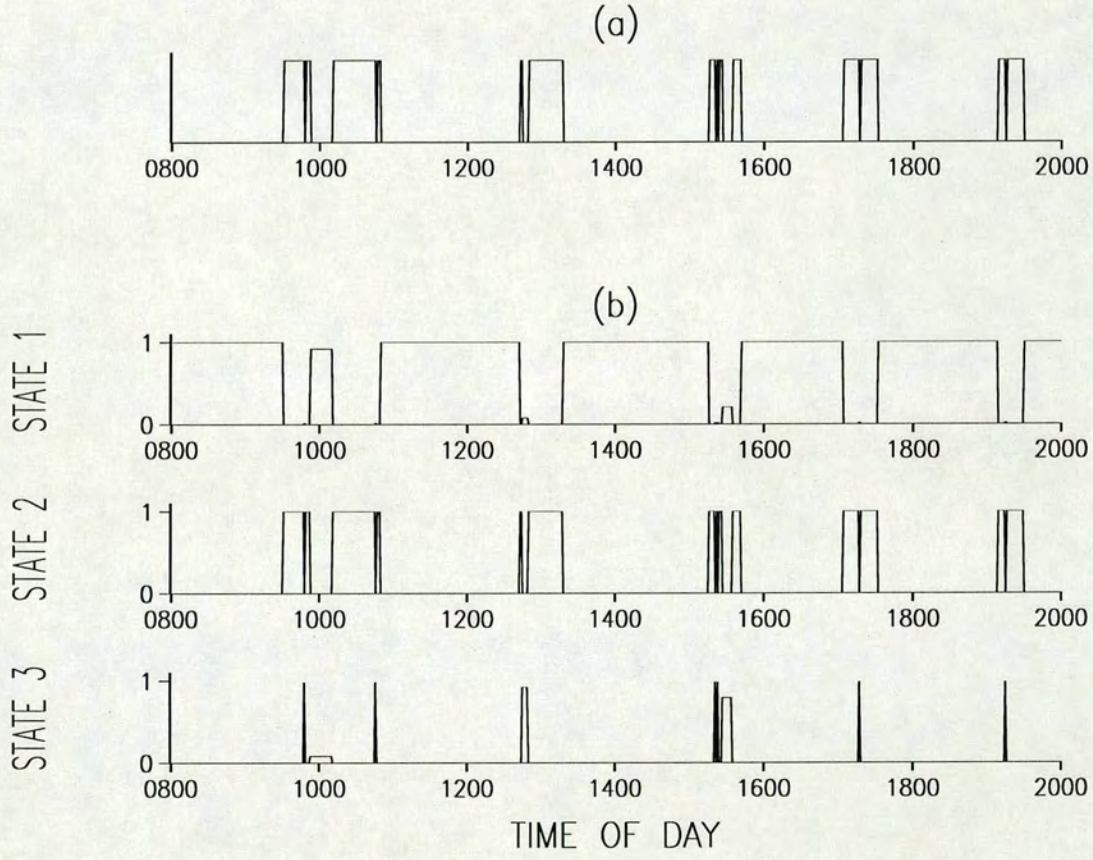

Figure 5.8: *Cow 108; (a) sample of data, (b) pointwise probabilities of being in each of the states for the three-state discrete-time compartment model shown in Figure 5.7.*

Cow	K	m	$-\mathcal{L}$	*AIC*	*BIC*
5	2	2	3792.07	7588.1	7605.5
	3	6	3350.91	6713.8	6765.9
	3	4	3350.91	**6709.8**	**6744.5**
41	2	2	4235.71	8475.4	8492.8
	3	6	3738.27	7488.5	7540.6
	3	4	3738.27	**7484.5**	**7519.2**
108	2	2	4441.16	8886.3	8903.7
	3	6	4002.52	8017.0	8069.1
	3	4	4002.52	**8013.0**	**8047.7**
169	2	2	3205.94	6415.9	6433.2
	3	6	2876.82	5765.6	5817.7
	3	4	2876.82	**5761.6**	**5796.3**
170	2	2	4685.14	9374.3	9391.6
	3	6	3974.94	7961.9	8013.9
	3	4	3974.94	**7957.8**	**7992.6**
182	2	2	3972.63	7949.3	7966.6
	3	6	3456.67	6925.3	6977.4
	3	4	3456.67	**6921.3**	**6956.0**
194	2	2	4165.69	8335.4	8352.7
	3	6	3719.13	7450.3	7502.3
	3	4	3719.13	**7446.3**	**7481.0**
221	2	2	6183.91	12371.8	12389.2
	3	6	5209.75	10431.5	10483.5
	3	4	5209.75	**10427.5**	**10462.2**

Table 5.7: *Log-likelihood and information criteria for two- and three-state discrete-time compartment models. K is the number of states and m the number of parameters. The preferred model for each cow according to each of the criteria is highlighted.*

Cow	*Model*	*AIC*	*BIC*
5	HMM	6716	6777
	Compartment	**6710**	**6745**
41	HMM	**7484**	7545
	Compartment	**7484**	**7519**
108	HMM	**7965**	**8023**
	Compartment	8013	8048
169	HMM	5767	5827
	Compartment	**5762**	**5796**
170	HMM	**7938**	7999
	Compartment	7958	**7993**
182	HMM	**6911**	6972
	Compartment	6921	**6956**
194	HMM	**7441**	7502
	Compartment	7446	**7481**
221	HMM	**10376**	**10436**
	Compartment	10428	10462

Table 5.8: *Comparison of three-state hidden Markov and discrete-time compartment models. The preferred model for each cow according to each of the criteria is highlighted.*

we have already pointed out, it is always BIC that prefers the more parsimonious model. In some cases where the criteria give similar values, it might also be the case that the simpler model is adequate.

5.3.6 Model diagnostics

We have seen that, depending on the individual animal, either the three-state hidden Markov or discrete-time compartment model appears to offer a satisfactory model for the cow feeding data, and we have seen from Figure 5.2 that data simulated from this model look similar to the observed data. At this stage, various model-checking procedures are available to further check the appropriateness of the model. One approach would be to compare the autocorrelation functions of the data and the fitted model, another would be to compare the expected marginal distributions of behaviour durations with those from the data. These and other ideas are outlined in MacDonald and Zucchini (1997, Sections 2.4, 2.6). Here we consider the marginal distributions of behaviours.

For simple Markov models in discrete time, the durations in a particular state follow a geometric distribution, which is the discrete analogue of the exponential distribution, and has density function

$$P(\text{duration} = \tau) = \lambda^{\tau-1}(1-\lambda) \quad \text{for} \quad \tau = 1, 2, 3, \ldots$$

For a hidden Markov model this can be generalised and we find that for a K-state model, the distribution of durations for any one behaviour is in general a mixture of K geometric distributions, the parameters of which are dependent on the transition probabilities Γ and the state-dependent probabilities π.

To see this, consider a K-state hidden Markov model with transition matrix Γ, with elements Γ_{kl} for $k, l = 1, \ldots, K$, which determines the overall stationary distribution of states, $\delta' = (\delta_1, \delta_2, \ldots, \delta_K)$. We also have the probabilities of feeding in each state, $(p_1, p_2, \ldots, p_K)$. Then, writing $\lambda(1) = \text{diag}(p_1, p_2, \ldots, p_K)$, we define the matrix $B = \Gamma\lambda(1)$. The probability of a particular behaviour being of duration τ is then a linear combination of $\lambda_k^{\tau-1}(1-\lambda_k), k = 1, \ldots, K$, where the λ_k are the eigenvalues of matrix B (MacDonald and Zucchini, 1997, pages 86–88). Hence the marginal distributions of durations of any behaviour are given by a mixture of geometric distributions.

This enables us to compute the marginal distributions for feeding and non-feeding durations in the models we have fitted. For the two-state HMM, the distribution

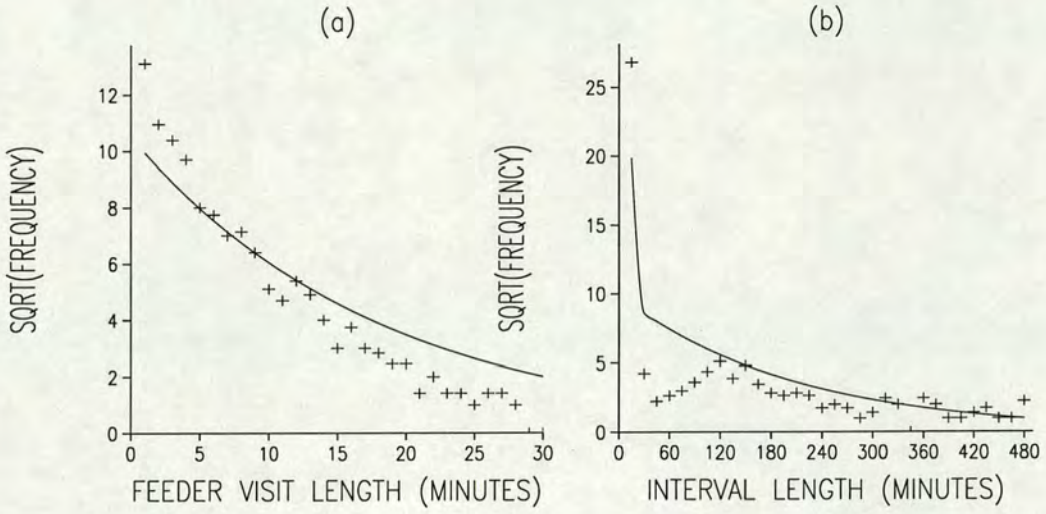

Figure 5.9: *Cow 108; marginal distributions of (a) feeder-visit lengths, (b) interval lengths; (+) sample frequencies, (—) frequencies based on the two-state hidden Markov model.*

of non-feeding durations is of the form

$$P(\text{duration} = \tau) = a\lambda_1^{\tau-1}(1-\lambda_1) + (1-a)\lambda_2^{\tau-1}(1-\lambda_2),$$

where λ_1 and λ_2 are the eigenvalues of the matrix B defined above and the mixing proportion a is computed from the eigenvalues and their associated eigenvectors.

For feeding durations the situation is simpler as there is zero probability of feeding in state 1. The distribution of feeding durations is therefore a simple geometric distribution with easily calculated parameters, given by

$$P(\text{duration} = \tau) = (\Gamma_{22}p_2)^{\tau-1}(1-\Gamma_{22}p_2).$$

Figure 5.9 shows these expected distributions superimposed on the observed histogram of feeding and non-feeding durations. We saw in Section 2.1.2 that an exponential distribution provided a reasonable description of feeding durations. Here, in a discrete time framework, the geometric is the corresponding distribution. However for non-feeding periods, we saw in Section 2.2 that a combination of exponential distributions could not adequately describe the shape of the distribution and that a mixture of log-normal distributions was preferable. We can see from Figure 5.9(b) that, as expected, the shape of the fitted distribution does not well describe the data, and although we illustrate here only with the two-state hidden Markov model, we have seen that even for a general K-state model, the duration of behaviours will still be constrained to follow a mixture of geometric distributions. Therefore, because a mixture of exponential functions is always

decreasing, the shape of the distribution of the data in Figure 5.9(b) will never be achievable with a HMM, no matter how many states it has. Other diagnostic techniques such as autocorrelation could also be investigated, but as we have already seen that this type of model cannot adequately describe the cow feeding data we shall not consider them here.

5.4 Summary

Hidden Markov models have an attractive interpretation biologically in that it is the underlying state of the animal that is being modelled by the Markov chain, then the observed behaviour is occurring conditional on this. It has been seen that a simple two-state model does not describe the cow feeding data as well as a three-state model. However even with this relatively small number of states there can be problems with parameter redundancy and so models with certain parameters for the transition probabilities between states were fixed. Formal statistical techniques to decide the optimal number of states are not generally applicable and so I have used information criteria, namely AIC and BIC, as a guide. In general, a three state model with no transition between the two states that were mainly non-feeding, was the best model of those considered. For some animals the Bernoulli probability of feeding in each of the states could be replaced by fixed feeding or non-feeding as appropriate, resulting in a simpler model whilst retaining as good a description of the data as the more general HMM. This simpler model was termed a discrete-time compartment model. However many states are used and whether general HMMs or compartment models are considered, we have seen that the durations in states are constrained to follow geometric distributions, and durations of particular behaviours follow mixtures of geometric distributions. Therefore unless this is appropriate for the dataset being considered, this model will always need to be generalised further.

Chapter 6

Semi-Markov models

In this chapter, the modelling of animal behaviour data using semi-Markov models is considered. After discussing the motivation for this in Section 6.1, I briefly outline the definitions and relationship with renewal processes in Section 6.2, also stating some basic results on Markov and semi-Markov chains. In Section 6.3, I look at the use of the EM algorithm in fitting the models, given that the state, when non-feeding, is unobserved. In Section 6.4 models are fit to the cow feeding data and comparisons made between models with different numbers of states. Finally, in Section 6.5, hidden semi-Markov models are briefly discussed.

6.1 Motivation

To motivate the use of a continuous-time semi-Markov model we note the following properties of the cow feeding data that we saw in Chapter 2.

- The marginal distribution of feeding durations can be described by an exponential distribution (Section 2.1.2).
- A mixture of two or three log-normal distributions is appropriate to describe the marginal distribution of non-feeding durations (Section 2.2.2).
- The durations of adjacent feeding events show no clear dependence (Figure 2.18). Therefore we assume them to be independent.
- There is some dependence between adjacent non-feeding events when classified as short or long (Sections 2.4.2 and 2.4.3).

These properties therefore suggest constructing a semi-Markov model for which non-feeding durations are dependent on preceding non-feeding durations, and these non-feeding periods are interspersed with feeding events, the durations of

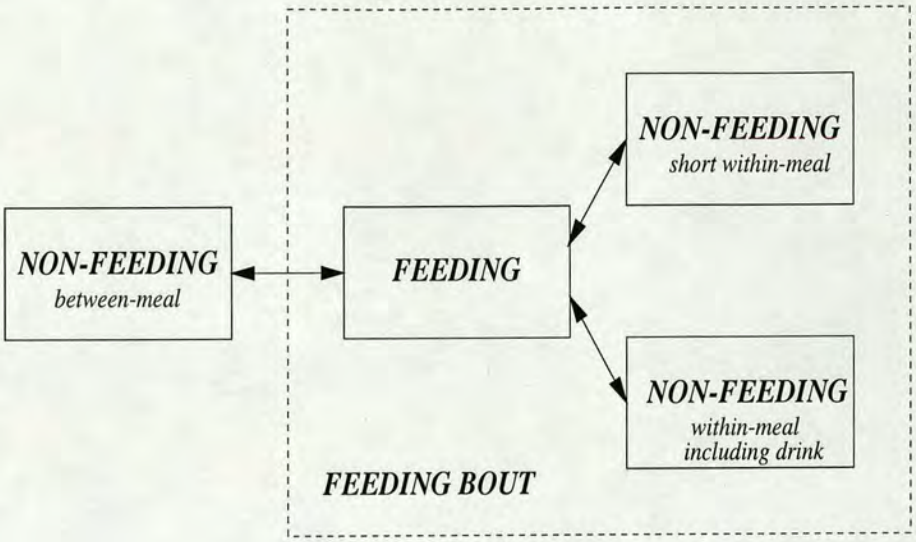

Figure 6.1: *Pictorial representation of a four-state semi-Markov model. There is one feeding state, one between-meal non-feeding state and two non-feeding states that are considered to be within a meal.*

which are independent. The semi-Markov model allows us to directly specify the marginal distributions of durations in each state, i.e. exponential for the feeding state and log-normal for each of two or three non-feeding states. We will consider three-state models, for which there is one feeding state and two non-feeding states, corresponding to short (within-meal) events or long (between-meal) events, and four-state models, for which the within-meal non-feeding category is further split into two, corresponding to short within-meal non-feeding periods and longer ones that contain drinking (see Section 2.2.2). The four-state model can be represented pictorially by Figure 6.1. Here the arrows show possible transitions between the states. For the three-state model we simply combine the within-meal non-feeding states, hence considering non-feeding periods within a meal to be described by a single distribution. Figure 6.2 shows a sample of data simulated from a three-state model, with parameters estimated from Cow 108. The simulated data can be seen to be similar to observed data. The states are ordered to allow easy comparison with Figure 5.2 in the last chapter, but the non-feeding states are numbered 1 and 2 to correspond with results presented in this chapter. State F is the feeding state.

6.2 Theory

There is often confusion as to the difference between types of semi-Markov and renewal process, partly because different authors have used the same names to

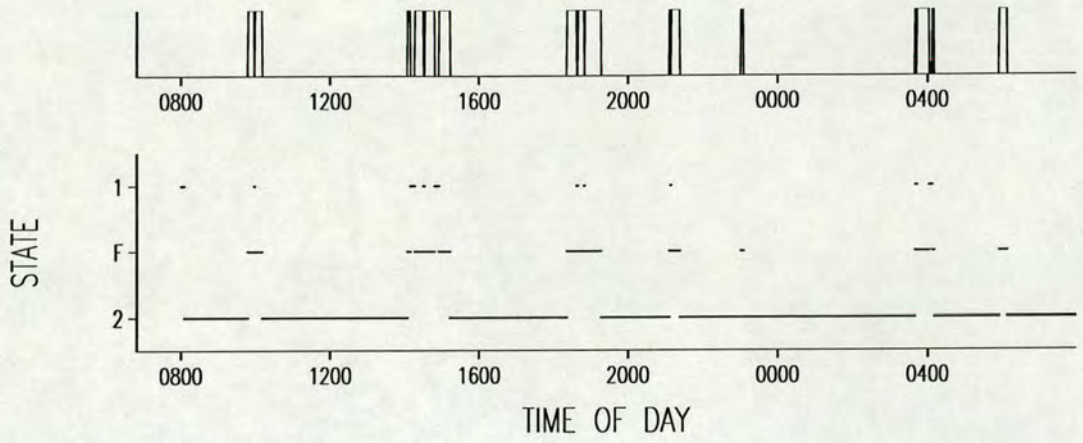

Figure 6.2: *Simulated data from a three-state semi-Markov model with parameters estimated from Cow 108. Lower graph shows current state: 1 and 2 are the two non-feeding states, F is the feeding state. Upper graph shows the corresponding feeding events that would be observed.*

denote different models. We defined these processes in Section 1.4 and here add some more details and provide some notation. We begin with the simplest sort of Markov process, the discrete-time Markov chain. We then consider how a continuous-time Markov chain relates to this and finally how to generalise further to the semi-Markov model. Semi-Markov processes are sometimes called *Markov-renewal processes* (Pyke, 1961), and if there are only two states are usually called *alternating renewal processes.* With only one state the process is simply a *renewal process.* Many of the details that follow are discussed in Haccou and Meelis (1994, Chapter 7) and can be found in other texts on Markov chains.

6.2.1 Discrete-time Markov chains

Consider a simple first-order Markov chain of length N, in discrete time on a finite state space $S = \{1, 2, \ldots, K\}$. We have a $K \times K$ matrix of transition probabilities Γ, with elements Γ_{kl}. The likelihood is of a multinomial form and can be easily maximised, giving the maximum likelihood estimate of the transition probability from state k to state l as

$$\hat{\Gamma}_{kl} = \frac{N_{kl}}{N_k}, \tag{6.1}$$

where N_{kl} is the total number of transitions from state k to l and $N_k = \sum_{l=1}^{K} N_{kl}$, the total number of observations in state k.

6.2.2 Continuous-time Markov chains

We want to develop models in continuous time and so now consider a continuous-time Markov chain with K states. Much of this theory can be found in Lehoczky (1998). The data take the form of a set of durations $\{\tau_i : i = 1, \ldots, N\}$ with an associated set of states $\{S_i : i = 1, \ldots, N\}$, where $S_i \in \{1, 2, \ldots, K\}$. For the discrete-time Markov chain there is a smallest unit of time between which transitions can occur, whereas in the continuous-time case, transitions can occur in arbitrarily small periods of time. For a time-homogeneous process, the probability of transition between two states depends only on the time in the current state and not on the overall time. Instead of working with a transition probability matrix, we now consider a transition rate matrix Ψ. Writing the transition probabilities as $\Gamma_{kl}(\tau)$, the elements of Ψ are defined by $\Psi_{kl} = \Gamma'_{kl}(0)$, i.e.

$$\Psi_{kl} = \begin{cases} \lim\limits_{h \to 0} \dfrac{\Gamma_{kk}(h) - 1}{h} & \text{if } k = l \\ \lim\limits_{h \to 0} \dfrac{\Gamma_{kl}(h)}{h} & \text{if } k \neq l. \end{cases}$$

The transition rate Ψ_{kl} can be interpreted as the probability per unit time of switching from state k to state l. In the discrete-time model we had

$$\sum_{l=1}^{K} \Gamma_{kl} = 1,$$

and now for the continuous case we have

$$\sum_{l=1}^{K} \Psi_{kl} = 0.$$

The durations in state k follow an exponential distribution, the maximum likelihood estimate of the rate parameter being

$$\hat{\lambda}_k = -\Psi_{kk} = \sum_{l \neq k} \Psi_{kl}.$$

The full likelihood for the process is given by

$$l = \left(\prod_{k=1}^{K} \prod_{l=1, l \neq k}^{K} \left(\frac{\Psi_{kl}}{-\Psi_{kk}} \right)^{N_{kl}} \right) \left(\prod_{k=1}^{K} \exp(\Psi_{kk} T_k) \right), \tag{6.2}$$

where T_k is the total time spent in state k, i.e.

$$T_k = \sum_{i=1}^{N} \delta_{ik} \tau_i,$$

where

$$\delta_{ik} = \begin{cases} 1 & \text{if } S_i = k \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

Hence sufficient statistics for the chain are the N_{kl} for $k, l = 1, \ldots, K$, together with the overall times spent in each state, T_k for $k = 1, \ldots, K$.

This gives maximum likelihood estimates of the transition rates as

$$\hat{\Psi}_{kl} = \frac{N_{kl}}{T_k}.$$

For the continuous-time Markov process, the sequence of states forms a discrete Markov chain. Approaching it in this way we can express the transition rate from k to l as the product of the termination rate of k and the transition probability of moving from state k to l, i.e.

$$\Psi_{kl} = \lambda_k \Gamma_{kl},$$

where λ_k is the exponential distribution parameter for durations in state k as described above and Γ_{kl} is the transition probability for moving from state k to l when considering the states as forming a discrete-time Markov chain. The maximum likelihood estimates for the transition probabilities are given by (6.1).

6.2.3 Semi-Markov chains

A semi-Markov process (Cox and Isham, 1980, page 54) is generalised from the Markov process simply by allowing specific distribution functions for the durations in each state, i.e. for state k we no longer have a constant termination rate λ_k and hence durations no longer follow an exponential distribution. To describe the cow feeding data, for non-feeding states we can replace the exponential distribution by a log-normal one, and the termination rate will now be the hazard function for the log-normal distribution.

A semi-Markov process can also be thought of as a Markov process for which the timescale has been transformed, or for which the transition probabilities are now a function of time. Therefore in fitting a semi-Markov model, the transition probabilities can be estimated by consideration of the states forming a discrete Markov chain as before, i.e. by (6.1). Then for termination rates, instead of the set of total times spent in each state forming the set of sufficient statistics as for the Markov process, we now generalise this to cover any marginal distribution for the durations in states. So for example, for a log-normal distribution, the set of

sufficient statistics would be the set of

$$\sum_{i=1}^{N} \delta_{ik} \log \tau_i \quad \text{and} \quad \sum_{i=1}^{N} \delta_{ik} (\log \tau_i)^2,$$

where δ_{ik} is given by (6.3) and τ_i is the duration in state S_i.

For fitting semi-Markov models to the cow feeding data there is a further complication in that when non-feeding, the current state of the animal is unknown. For the three-state model, non-feeding corresponds to either of two states; for the four-state model there are three non-feeding states. Therefore the approach we take is to treat the non-feeding states as missing and estimate parameters using the EM algorithm.

6.3 Fitting a semi-Markov model using the EM algorithm

Consider for a start the three-state model. Non-feeding events could be classified according to the meal criteria derived in Section 2.2 and the methods described in the previous section used to estimate transition probabilities and termination rates. However if we work directly with such classifications, we will be ignoring the fact that some events have been classified incorrectly, which will affect the parameter estimates. In our case, distributions were well-separated and hence problems with misclassification were small, but for situations in which the distributions are more overlapping, this misclassification should be allowed for. Therefore we consider a method that uses the EM algorithm to estimate parameters, treating the states as missing. We consider first-order Markov models only, but the methodology is easily extendable to a higher order Markov chain and the inclusion of covariates, e.g. diurnal effect. Note that as we are assuming feeding durations to be independent both of each other and of the non-feeding durations, we fit the semi-Markov model only to the non-feeding states.

We have already described the EM algorithm in the form of the Baum-Welch algorithm for hidden Markov models in Chapter 5. The EM algorithm itself was first so-called by Dempster et al. (1977) and is an iterative scheme which can be used when estimating parameters for a model which is being fit to incomplete data. For our application, the complete data can be written as the set of $(S_i, \tau_i), i = 1, \ldots, N$, where the S_i are the set of non-feeding states and the τ_i are the durations in each state. The S_i are unobserved and so form the missing data.

We have the conditional densities

$$\log \tau_i | (S_i = k) \sim N(\mu_k, \sigma_k^2) \text{ for } k = 1, \ldots, K, \tag{6.4}$$

where K is the number of non-feeding states. To start the algorithm we need initial estimates of the parameters — $\mu_k, \sigma_k^2, k = 1, \ldots, K$, plus a further $K(K-1)$ parameters for the transition probabilities. We then carry out alternate steps of calculating the sufficient statistics of the data given the current parameter estimates, and maximising the likelihood given these sufficient statistics and the current parameter estimates. This is repeated until convergence in the parameter estimates is achieved.

In detail, we have the following.

- Begin with initial estimates of the parameters, $\mu_k, \sigma_k^2, k = 1, \ldots, K$ plus $K(K-1)$ parameters for transition probabilities, so we could take

  $$\Gamma_{kl} = P(S_i = l | S_{i-1} = k)$$

  for $k = 1, \ldots, K$ and $l = 1, \ldots, K-1$, the Γ_{kk}'s being determined by the constraint that the sum of each row of the matrix must sum to 1. These also in turn determine $\delta' = (\delta_1, \ldots, \delta_K)$, the stationary probabilities of the states, via the equation $\delta' = \Gamma\delta'$ subject to the constraint $\sum_{k=1}^{K} \delta_k = 1$.

- Step 1 — Expectation.

  We need to work out the expectation of the sufficient statistics of the complete data, given the observed data and the current parameter estimates. We can do this via the forward-backward algorithm described in Lindgren (1978) and Le et al. (1992), which we discussed in detail in Section 5.2.2. There, we were working with a minutely series (indexed $t = 1, \ldots, n$) of feeding/non-feeding observations (x_t) and a set of unobserved states (C_t). Here we are working with a series of events (indexed $i = 1, \ldots, N$) of observed duration (τ_i) but associated with an unobserved state (S_i). Since the situation and notation are now different we present the formulae again here.

  The forward and backward probabilities are α_{ik}, the joint probability of the observed non-feeding durations up to the i-th event, and the i-th state being k, and β_{ik}, the conditional probability of the observed set of durations from $i+1$ onwards, given that the i-th state is k, i.e.

  $$\begin{aligned} \alpha_{ik} &= P(\tau_1, \ldots, \tau_i, S_i = k) \\ \beta_{ik} &= P(\tau_{i+1}, \ldots, \tau_N | S_i = k), \end{aligned}$$

for events $i = 1, \ldots, N$ and states $k = 1, \ldots, K$.

These can be calculated from

$$\begin{aligned}
\alpha_{1k} &= P(S_1 = k)\, P(\tau_1 | S_1 = k) \\
&= \delta_k\, P(\tau_1 | S_1 = k), \\
\alpha_{ik} &= \sum_{l=1}^{K} \alpha_{i-1,l}\, \Gamma_{lk}\, P(\tau_i | S_i = k) \quad \text{for} \quad i = 2, \ldots, N, \\
\beta_{Nk} &= 1, \\
\beta_{ik} &= \sum_{l=1}^{K} \beta_{i+1,l}\, \Gamma_{kl}\, P(\tau_{i+1} | S_{i+1} = l) \quad \text{for} \quad i = N-1, N-2, \ldots, 1.
\end{aligned}$$

In order to obtain all the sufficient statistics of the data we need to calculate the probabilities of the current state and of pairs of states, conditional on the observed durations, $\tau_1, \ldots, \tau_N$. These are respectively given by γ_{ik}, the probability of the i-th state being k, and η_{ikl}, the probability of the i-th state being k and the $(i+1)$-th being l, conditional on the whole observed series of durations. The γ_{ik} are given by

$$\begin{aligned}
\gamma_{ik} &= P(S_i = k | \tau_1, \ldots, \tau_N) \\
&= \frac{P(\tau_1, \ldots, \tau_i, S_i = k)\, P(\tau_{i+1}, \ldots, \tau_N | S_i = k)}{\sum_{k=1}^{K} P(\tau_1, \ldots, \tau_i, S_i = k)\, P(\tau_{i+1}, \ldots, \tau_N | S_i = k)} \\
&= \frac{\alpha_{ik}\, \beta_{ik}}{\sum_{k=1}^{K} \alpha_{ik}\, \beta_{ik}},
\end{aligned}$$

for $i = 1, \ldots, N$ and $k = 1, \ldots, K$, equivalent to (5.5).

The η_{ikl} are given by

$$\begin{aligned}
\eta_{ikl} &= P(S_i = k, S_{i+1} = l | \tau_1, \ldots, \tau_N) \\
&= \frac{P(S_i = k | \tau_1, \ldots, \tau_N)\, \Gamma_{kl}\, P(\tau_{i+1} | S_{i+1} = l)\, P(\tau_{i+2}, \ldots, \tau_N | S_{i+1} = l)}{P(\tau_{i+1}, \ldots, \tau_N | S_i = k)} \\
&= \frac{\gamma_{ik}\, \Gamma_{kl}\, P(\tau_{i+1} | S_{i+1} = l)\, \beta_{i+1,l}}{\beta_{ik}}
\end{aligned}$$

for $i = 1, \ldots, N-1$ and $k, l = 1, \ldots, K$.

We now have the conditional probabilities required for the sufficient statistics of the complete data. Note also that the current value of the log-likelihood can be calculated as $\mathcal{L} = \log \sum_{k=1}^{K} \alpha_{Nk}$.

- Step 2 — Maximisation.

  Using the expectation of the sufficient statistics of the complete data, all parameters are re-estimated by full maximum likelihood, using the equations

$$\begin{aligned}
\hat{\mu}_k &= \frac{\sum_{i=1}^{N} \gamma_{ik} \log \tau_i}{\sum_{i=1}^{N} \gamma_{ik}}, \\
\widehat{\sigma_k^2} &= \frac{\sum_{i=1}^{N} \gamma_{ik} (\log \tau_i - \hat{\mu}_k)^2}{\sum_{i=1}^{N} \gamma_{ik}}, \\
\hat{\Gamma}_{kl} &= \frac{\sum_{i=1}^{N} \eta_{ikl}}{\sum_{l=1}^{K} \left[\sum_{i=1}^{N} \eta_{ikl} \right]}, \\
\hat{\delta}_k &= \frac{1}{N} \sum_{i=1}^{N} \gamma_{ik}.
\end{aligned}$$

- We now return to step 1, using the new parameter estimates. Steps 1 and 2 are repeated until convergence in the parameter estimates is achieved.

Hence this algorithm provides parameter estimates which take into account the uncertainty associated with classifying intervals as within- or between-meal and also gives estimates of the transition probabilities for the first-order Markov sequence of states.

Diurnal effect can be allowed for by modification of transition probabilities by an approach analogous to that used for hidden Markov models in Section 5.2.3. However we have not investigated this here, again because of the lack of a suitable parametric form to adequately describe the diurnal effect. In addition the methodology could be extended to higher order Markov chains, i.e. the current non-feeding duration dependent on the previous r non-feeding durations, in exactly the same way as a first order Markov model is generalised to a r-th order one. However for simplicity here we illustrate with just the first order model.

6.4 Fitting to cow feeding data

Results are presented for the models fit to the eight high-protein cows. We first look at three-state models, with two states for non-feeding, interpreted simply as within- and between-meal. We also look at four-state models, thus splitting the within-meal non-feeding periods into those which include a drink and those that do not, as discussed in Section 2.2.2. Feeding events play no part in the model fitting here, assumed to be independent and exponentially distributed. A composite log-likelihood could be formed by the addition of maximised log-likelihoods for the feeding events as in Section 2.2.2, and the non-feeding durations as in this section.

6.4.1 Three-state models

Table 6.1 shows parameter estimates for the eight high-protein cows, for three-state semi-Markov models that have one state for feeding and two states for non-feeding. Estimates of μ_k, σ_k^2, for $k = 1, 2$, and the overall (stationary) probabilities of the states can be compared with those from Table 2.4, when essentially the same model could be fit in stages — fitting the mixture of distributions, determining a meal criterion, classifying the intervals and lastly estimating transition probabilities. In contrast, the method using the EM algorithm estimates all parameters simultaneously. Although differences in results between the methods are negligible in this case, for distributions that are more overlapping and hence for which a meal criterion is less well-defined, the second method is much more satisfactory.

For feeding events we are assuming independence and hence simply have to estimate the parameter for the exponential distribution describing the durations. This was done in Section 2.1.2 and parameter estimates are as given in Table 2.2.

6.4.2 Four-state models

Table 6.2 shows parameter estimates for the three non-feeding states of four-state semi-Markov models. Comparison with Table 2.5 shows that some of the estimates are now quite different to those obtained when the model was fit in stages assuming that there was no error in classification of intervals into the three types. For cows that display a clear third distribution, e.g. Cow 108, differences between the estimates are negligible, whereas more discrepancy is seen for animals

Cow	$\hat{\Gamma}$	$\hat{\delta}'$	$(\hat{\mu}_1, \hat{\mu}_2)$ $(\hat{\sigma}_1, \hat{\sigma}_2)$
5	$\begin{pmatrix} .6354 & .3646 \\ .8549 & .1451 \end{pmatrix}$	(.7012, .2988)	(4.309, 9.349) (1.305, 0.473)
41	$\begin{pmatrix} .7483 & .2517 \\ .9343 & .0657 \end{pmatrix}$	(.7867, .2133)	(4.469, 9.461) (1.472, 0.426)
108	$\begin{pmatrix} .7498 & .2502 \\ .9121 & .0879 \end{pmatrix}$	(.7850, .2150)	(3.787, 9.112) (1.416, 0.573)
169	$\begin{pmatrix} .5878 & .4122 \\ .7878 & .2122 \end{pmatrix}$	(.6572, .3428)	(4.208, 9.320) (1.226, 0.634)
170	$\begin{pmatrix} .7669 & .2331 \\ .9396 & .0604 \end{pmatrix}$	(.8015, .1985)	(4.123, 9.276) (1.205, 0.577)
182	$\begin{pmatrix} .7897 & .2103 \\ .9193 & .0807 \end{pmatrix}$	(.8141, .1859)	(4.413, 9.609) (1.478, 0.533)
194	$\begin{pmatrix} .6570 & .3430 \\ .8656 & .1344 \end{pmatrix}$	(.7166, .2834)	(3.796, 9.043) (1.324, 0.653)
221	$\begin{pmatrix} .8200 & .1800 \\ .9603 & .0397 \end{pmatrix}$	(.8422, .1578)	(3.867, 9.004) (1.240, 0.624)

Table 6.1: *Parameter estimates for the two non-feeding states of three-state semi-Markov models.* $\hat{\Gamma}$ *is the transition probability matrix and* $\hat{\delta}$ *the vector of stationary probabilities for the non-feeding states.*

Cow	$\hat{\Gamma}$	$\hat{\delta}'$	$(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3))$ $(\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3))$
5	$\begin{pmatrix} .3068 & .4351 & .2582 \\ .1853 & .3634 & .4514 \\ .6470 & .2201 & .1330 \end{pmatrix}$	(.3627, .3482, .2890)	(3.787, 4.967, 9.380) (0.921, 1.521, 0.440)
41	$\begin{pmatrix} .2425 & .7058 & .0517 \\ .0963 & .6040 & .2997 \\ .2766 & .6595 & .0639 \end{pmatrix}$	(.1571, .6310, .2120)	(3.133, 4.811, 9.466) (0.592, 1.446, 0.423)
108	$\begin{pmatrix} .5821 & .1575 & .2604 \\ .6488 & .1336 & .2176 \\ .7558 & .1506 & .0936 \end{pmatrix}$	(.6305, .1522, .2173)	(3.234, 6.015, 9.101) (0.915, 0.596, 0.581)
169	$\begin{pmatrix} .3289 & .3088 & .3623 \\ .2623 & .2883 & .4494 \\ .4483 & .3477 & .2040 \end{pmatrix}$	(.3475, .3166, .3359)	(3.442, 5.128, 9.350) (0.694, 1.181, 0.600)
170	$\begin{pmatrix} .4922 & .4432 & .0645 \\ .0000 & .7036 & .2964 \\ .5584 & .3802 & .0614 \end{pmatrix}$	(.2190, .5823, .1986)	(3.371, 4.405, 9.276) (0.887, 1.187, 0.578)
182	$\begin{pmatrix} .2404 & .5606 & .1990 \\ .2822 & .5134 & .2044 \\ .3706 & .5583 & .0712 \end{pmatrix}$	(.2865, .5348, .1787)	(3.232, 5.102, 9.650) (0.643, 1.455, 0.492)
194	$\begin{pmatrix} .6182 & .0879 & .2939 \\ .0423 & .5695 & .3882 \\ .4642 & .4113 & .1245 \end{pmatrix}$	(.3790, .3425, .2786)	(3.127, 4.595, 9.062) (0.911, 1.365, 0.638)
221	$\begin{pmatrix} .7370 & .0743 & .1887 \\ .8051 & .0988 & .0960 \\ .8853 & .0733 & .0414 \end{pmatrix}$	(.7652, .0766, .1582)	(3.628, 6.238, 9.003) (1.016, 0.522, 0.621)

Table 6.2: *Parameter estimates for the three non-feeding states of four-state semi-Markov models.* $\hat{\Gamma}$ *is the transition probability matrix and* $\hat{\delta}$ *the vector of stationary probabilities for the states.*

Cow	*N*	*K*	m_K	-$\mathcal{L}$	*AIC*	*BIC*
5	587	3	6	1147.26	2306.5	**2332.9**
		4	12	1132.40	**2288.8**	2341.3
41	730	3	6	1476.68	2965.4	**2993.0**
		4	12	1458.12	**2940.2**	2995.3
108	944	3	6	1951.60	3915.2	3944.3
		4	12	1889.37	**3802.7**	**3861.0**
169	504	3	6	1009.80	2031.6	**2056.9**
		4	12	992.48	**2009.0**	2059.7
170	897	3	6	1733.43	3478.9	3507.6
		4	12	1705.54	**3435.1**	**3492.6**
182	683	3	6	1414.31	2840.6	2867.8
		4	12	1386.14	**2796.3**	**2850.5**
194	771	3	6	1588.37	3188.9	3216.7
		4	12	1560.43	**3144.7**	**3200.6**
221	1323	3	6	2567.45	5146.9	5177.9
		4	12	2534.11	**5092.2**	**5154.5**

Table 6.3: *Log-likelihoods and information criteria for three- and four-state semi-Markov models. N is the number of events recorded, m_K the number of parameters for the K-state model and $\mathcal{L}$ the maximised log-likelihood.*

that had a less well-defined third distribution, e.g. Cow 5. Comparison of the stationary probabilities in this table and Table 6.1 again shows that the overall classification of intervals into the between-meal state is fairly consistent between both models. As expected, the extra state in the four-state model results in the within-meal state from the three-state model being split into two.

6.4.3 Comparison of three- and four-state models

Table 6.3 shows values for the log-likelihood for the models fit to non-feeding times. In addition, the values for AIC and BIC are shown. For all cows the value of AIC is lower for the four-state model, in spite of the extra parameters that need to be estimated. However for three of the cows it can be seen that BIC points to the three-state model being preferable, and for some other cows there is not much difference between them. The biggest difference in favour of the four-state model is for Cow 108, for which we saw from Figure 2.9 there is strong

evidence for three non-feeding states. For Cow 170, illustrated in Figure 2.10, BIC points to the four-state model being slightly better, and from the figure it can be seen to be debatable whether a mixture of two distributions is adequate to describe non-feeding durations, or whether a third is necessary. For Cow 5, Figure 2.8 showed no evidence at all of a third distribution and this is confirmed by BIC being lower for the three-state model. So it seems that here, the more parsimonious BIC is more useful for model selection that AIC.

6.4.4 Latent states

We have seen that semi-Markov models appear to offer a good description of the cow feeding data. In terms of marginal distributions at least, these models are to be preferred over the hidden Markov models of the previous chapter. However here we do not have the extra layer of stochasticity in terms of an animal remaining in a given state whilst displaying different behaviours. However we do have this implicitly, in that it is easy to think about creating a new set of states for which if the cow is in either the feeding state or one of the within-meal non-feeding states, she is in a 'meal' state, and if she is in the between-meal non-feeding state, this is simply a 'between-meal' state.

The models in this chapter have been fit using the EM algorithm and there is much connection with the methodology of Chapter 5 for fitting HMMs. Therefore here also the Viterbi algorithm or calculation of the pointwise probabilities of being in each state can be applied, as described in Section 5.2.5. Classification of which non-feeding periods lie within a meal and which are between meals can then be carried out, so in Figure 6.2 we can classify periods entirely within states 1 and F as meals, and periods in state 2 as between-meal.

This model then has all the desired properties, although we may not have allowed for serial dependence and dependence on covariates sufficiently; the model could be extended to have higher order Markov dependence and also to incorporate diurnal effect in some way. These issues have already been discussed and are simply pointed out here as refinements that might be made to such models for future work and fitting to other datasets.

6.5 Hidden semi-Markov models

A hidden semi-Markov model is a generalisation of the hidden Markov model for which the Markov property is relaxed is the same way as from going from an ordinary Markov model to a semi-Markov one. Rabiner (1989) addresses what changes need to be made to the HMM methodology. Diagonal entries in the transition probability matrix are set to zero, as there cannot be transitions from a state back to itself in a model which specifically includes the state duration. Details on the computations involved in fitting this type of model are given in Sansom and Thomson (2000) and are seen to be considerably more intensive than for the HMM. Although HSMMs seem to be philosophically appropriate models for animal behaviour, some initial work with them suggested that the amount of computation involved would be prohibitive to them being suitable for general use. It is also difficult to imagine them offering any real advantage over the semi-Markov models already explored in this chapter.

6.6 Summary

In this chapter we have seen that semi-Markov models are capable of capturing the main features of the data, both in terms of the marginal distributions of feeding and non-feeding periods, and the dependence of the non-feeding periods and independence of feeding periods. I reviewed the main points on model fitting for continuous-time Markov chains and summarised modifications for the fitting of semi-Markov models. Further, to fit models to the cow feeding data, it must be taken into account that the set of states of non-feeding periods is unknown. Therefore the data was treated as incomplete — known durations but unknown states — and the EM algorithm was used to estimate all parameters. Assuming independence of feeding events, a composite likelihood can include a component from the marginal distribution of feeding events. For the cow feeding data it was seen that, as expected, the methodology using the EM algorithm here is superior to the ad-hoc approach used in Chapter 2 when marginal distributions were fitted to non-feeding durations, non-feeding durations classified into states using meal criteria and then transition probabilities could have been estimated for a first-order Markov model. Here the method estimates all parameters simultaneously and the uncertainty about the current non-feeding state is built into the estimation procedure. Information criteria were used to show that for some cows, two states are adequate for describing the non-feeding periods, whereas other cows

benefit from a third state. The use of hidden semi-Markov models was briefly discussed, but subsequently dismissed due to the intensity of computation needed for model fitting and hence its lack of suitability for general modelling of behaviour.

Chapter 7

Model comparisons

In preceding chapters, three main classes of model have been considered — the latent Gaussian models of Chapter 3, the hidden Markov models of Chapter 5 and the semi-Markov models of Chapter 6. I have considered the motivation for each, techniques for parameter estimation and some basic ideas of model choice and validation. In Section 7.1, I now consider the three types of model simultaneously and discuss their relative merits and the connections between them. After reviewing some of the literature on the comparison of non-nested models in Section 7.2, Section 7.3 goes on to apply a parametric bootstrap approach to the three types of model fit to the cow feeding data. Details of simulations are presented and I assess which type of model is most suitable for the cow feeding dataset.

7.1 Summary of models

For each cow we have considered at least six models, falling into the three categories as described above.

- Latent Gaussian models.

  Assuming the data occur from the thresholding of an underlying Gaussian process, the autocorrelation of either the observed binary process or the continuous latent process can be estimated and then this estimate of the correlation structure used to fit an ARMA model. An ARMA(2,1) process was seen to be the most parsimonious to describe the observed autocorrelation, and use of either binary or Gaussian correlation gave similar results if a sufficiently high number of lags were used.

- Hidden Markov models.

We considered models with two or three states. When including three states we found that in the interests of parsimony we could set the transition probabilities between the two mainly non-feeding states to zero without any decrease in likelihood. We also considered setting the state-dependent probabilities to 0 or 1, resulting in a model which we called a discrete-time compartment model, equivalent to a discrete-time Markov chain model. For some cows this model is as good a description as the three-state HMM, for others the HMM is preferable.

- Semi-Markov models.

  Feeding durations were considered to be independent, with marginal distribution well-described by exponential distributions. The marginal distribution of non-feeding periods could be described as a mixture of two or three log-normal distributions; the model also incorporates first-order dependency in types of non-feeding period.

7.1.1 Continuous or discrete time

I have already discussed how the data occur naturally in continuous time and are recorded to the nearest second, yet some of the models considered exist only in discrete time. A minutely discretisation scale has been used throughout, however in theory we could have used data to the nearest second. This was not done for two reasons, firstly to keep series lengths more manageable for computational reasons, and secondly, when the discretisation unit decreases, parameter estimates typically get nearer to the boundaries of their parameter space and so problems can be encountered in estimation procedures that involve numerical minimisation routines. For example, with the latent Gaussian model, we showed in Section 3.7.2.1 that the choice of an ARMA(2,1) model ensured that under a change of timescale the same model is retained. However, with parameter estimates close to the boundary even for the minutely timescale, problems could be anticipated if the timescale was reduced to seconds.

From a philosophical point of view, it is desirable for the model to at least exist in continuous time. The semi-Markov model occurs in continuous time anyway. In Section 3.7.2.2 we discussed how the latent Gaussian model does have a continuous time analogue, but that with the ARMA(2,1) class of model used, there are problems with it not being locally smooth. However it is thought that by using a model with more parameters, e.g. ARMA(3,2) these constraints could be imposed, although the details of this have not been fully investigated. The hidden

Markov models do not have a continuous-time analogue. Semi-Markov models therefore have the advantage over the others of existing naturally in continuous time. However we have already commented that if the discretisation unit is small enough to capture all the observed changes in behaviour then a discrete-time model may not necessarily be inferior to one in continuous time.

7.1.2 Latent structure

All the models involve some latent structure but it is incorporated in different ways. The different models may therefore have particular relevance to certain applications. The hidden Markov model has the obvious attraction of allowing the underlying state to remain unchanged, whilst different behaviours are being performed. This is easily extendable to large numbers of behaviours. In the case of feeding data, or any other situation in which we are concerned only with the occurrence of a single behaviour, maybe a more natural model is one which has a continuously changing latent variable. This would not be an obvious choice of model if multiple behaviours were being considered, although some options are discussed in Chapter 8. The semi-Markov model also has an implicit latent structure in that when the cow is not feeding we do not know which of the non-feeding states it is in and so whether the period can be considered within- or between-meal. We have seen that the same techniques can be applied as to hidden Markov models to decide either on the most likely sequence of states or to obtain pointwise probabilities for being in each state.

7.1.3 Diurnal variation and serial dependency

Diurnal patterns can be built into all of the models considered. For the latent Gaussian model, instead of having a fixed threshold over the whole day, we allowed it to depend on the overall probability of feeding at that time of day using a moving average approach. For hidden Markov models time can be built in as a covariate at the level of either the transition probabilities or the state-dependent probabilities of feeding; similarly for semi-Markov models the transition probabilities can be modelled in terms of time. However for the cow feeding, the extent of diurnal variation is very variable between animals, some displaying strong patterns, others looking more random. Hence we could not find a satisfactory parametric form to describe the effect and so resorted to using distinct values on an hourly basis. This generally results in a large increase in the number of parameters to be estimated relative to the size of the increase in likelihood achieved.

Nevertheless we have at least addressed how time-trend may be included in each of the models.

Serial dependency is implicit in all of the models considered. For the latent Gaussian model it is the continuous latent variable, an ARMA process, that contains all the information on dependency. For the Markov models we fit models with only first-order dependency, but methodology extends easily to higher-order Markov chains.

7.2 Comparison of non-nested models

For the comparison of nested models, the likelihood ratio test may be used to assess the relative fit of the two models, twice the difference in log-likelihood having an asymptotic chi-squared distribution, with degrees of freedom equal to the difference in number of parameters estimated. However when models are separate, i.e. non-nested, the theory on which the likelihood ratio test is based does not hold. Information criteria such as AIC or BIC are available, but these only give a numerical comparison and no level of significance can be attached. In our case even these criteria are not obviously applicable, because of the different ways in which the models are fit. For example the semi-Markov model is considered within a continuous-time framework, the others being fit to discretised data. Further, parameters for the latent Gaussian model are estimated by least squares, and for the others a likelihood is maximised.

In this section we first review methodology that has been developed in the literature for the comparison of separate models, and then we go on to develop a parametric bootstrap approach which can be applied to the models fit to the cow feeding data.

7.2.1 Cox statistics

Cox (1961, 1962) initiated work on the problem of comparing non-nested models, with a modification of the Neyman-Pearson likelihood ratio. Using notation similar to Cox's, the basic idea is to consider two models, H_1 with parameters α, and H_2 with parameters β. Letting $\mathcal{L}_1(\hat{\alpha})$ be the maximised log-likelihood under H_1 and $\mathcal{L}_2(\hat{\beta})$ that under H_2, we define

$$\mathcal{L}_{12} = \mathcal{L}_1(\hat{\alpha}) - \mathcal{L}_2(\hat{\beta}).$$

		T_1		
		$-$	NS	$+$
T_2	$-$	ev against H_1		ev against both
	NS	towards H_2	no ev against either	
	$+$	ev against both	ev against H_2 towards H_1	

Table 7.1: *Possible conclusions to draw from the Cox statistics T_1 and T_2. A result in one of the empty cells would indicate contradictory evidence (ev) from the two statistics. (NS indicates not significant.)*

For simple discrimination between two models we may simply want to consider this quantity, but to think in terms of a hypothesis testing situation, we can take the maximum likelihood estimators under one of the models and consider their distributions under the other model, via the statistics T_1 and T_2, defined by

$$\begin{aligned} T_1 &= \mathcal{L}_{12} - E_{\hat{\alpha}}[\mathcal{L}_{12}] \\ &= [\mathcal{L}_1(\hat{\alpha}) - \mathcal{L}_2(\hat{\beta})] - [\mathcal{L}_1(\hat{\alpha}) - \mathcal{L}_2(\beta_{\hat{\alpha}})] \\ T_2 &= \mathcal{L}_{12} - E_{\hat{\beta}}[\mathcal{L}_{12}] \\ &= [\mathcal{L}_1(\hat{\alpha}) - \mathcal{L}_2(\hat{\beta})] - [\mathcal{L}_1(\alpha_{\hat{\beta}}) - \mathcal{L}_2(\hat{\beta})], \end{aligned}$$

where notation $\beta_{\hat{\alpha}}$ is used to indicate the value of β under the maximum likelihood estimate of α, i.e. the value of β given H_1 is the true model.

T_1 compares $\mathcal{L}_{12}$ with the best estimate of the value it is expected to take under H_1, i.e. it is $\mathcal{L}_{12}$ minus its expectation under the null hypothesis. If H_1 was nested within H_2, then $-2\mathcal{L}_{12}$ would have a chi-squared distribution. However for non-nested models, $\mathcal{L}_{12}$ can be of either sign and, as might be expected, and as shown by Cox, these test statistics are asymptotically normal under H_1 and H_2 respectively. A large negative value of T_1 provides evidence against H_1 in the direction of H_2, and, analogously, a large positive value of T_2 is evidence of a departure from H_2 in the direction of H_1. Hence consideration of the two together will conclude one of four possibilities, either the data are consistent with both models, neither model, or with one but not the other. Table 7.1 shows the possible outcomes. Due to the form of T_1 and T_2 the table is antisymmetric. Cells that are empty would indicate contradictory evidence, e.g. a result in the middle cell of the top row would indicate no evidence from T_1 against model H_1, whereas T_2 indicates evidence against model H_2, but in the direction away from

Model	*Log Normal*		*Expl*	$\mathcal{L}_1$	$\mathcal{L}_2$	T_1^*	T_2^*
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\beta}$				
$\log X \sim \mathrm{N}(4, 0.75^2)$	4.00	0.76	72.94	−5147	−5290	−0.30	7.96
$X \sim \mathrm{Exp}(1/70)$	3.64	1.32	68.58	−5333	−5228	−7.35	−0.49
$X \sim \mathrm{Gamma}(2, 40)$	4.02	0.79	72.45	−5201	−5283	−6.81	5.88

Table 7.2: *Maximum likelihood estimates, maximised log-likelihood values,* $\mathcal{L}_1$ *and* $\mathcal{L}_2$*, and standardised Cox statistics,* T_1^* *and* T_2^**, for both the log-normal,* H_1*, and exponential,* H_2*, models, for data simulated under the three models shown.*

H_1. Apart from the two empty cells, the other cells all represent consistent results from the two statistics, either no evidence against either, evidence against both, or evidence in favour of one of the two models.

7.2.1.1 Example from Cox (1961)

We illustrate the use of the Cox statistics with an example given in Cox (1961), which compares the fit of exponential and log-normal distributions. The hypotheses to be tested are

$$H_1 : \log X \sim \mathrm{N}(\mu, \sigma^2)$$

against

$$H_2 : X \sim \mathrm{Exp}(1/\beta).$$

We illustrate with three situations. Firstly we simulate data from a log-normal distribution with mean 4 and variance 0.75^2, secondly data from an exponential distribution with mean 70, and thirdly data from a different distribution, a gamma with shape parameter 2 and scale parameter 40. Parameter values were chosen to give distributions with similar means.

Explicit forms for the Cox statistics for this situation are given in Cox (1961, equations 60, 71) and we denote by T_1^* and T_2^* the standardised values of the statistics, simply T_1 and T_2 respectively divided by the square root of their asymptotic variances, hence these have asymptotic standard normal distributions. The possible conclusions are given by Table 7.1. Table 7.2 shows the results for the simulated data, consisting of 1000 data points simulated from log-normal, exponential and gamma distributions, and Figure 7.1 shows histograms of these simulated datasets along with the fitted distributions. From this figure it can be seen that the correct distribution is a good fit to the simulated data for (a) and (b), whereas for (c) neither is a good fit, although the log-normal is seen to fit more closely than the exponential. The standardised Cox statistics in Table 7.2 confirm the following.

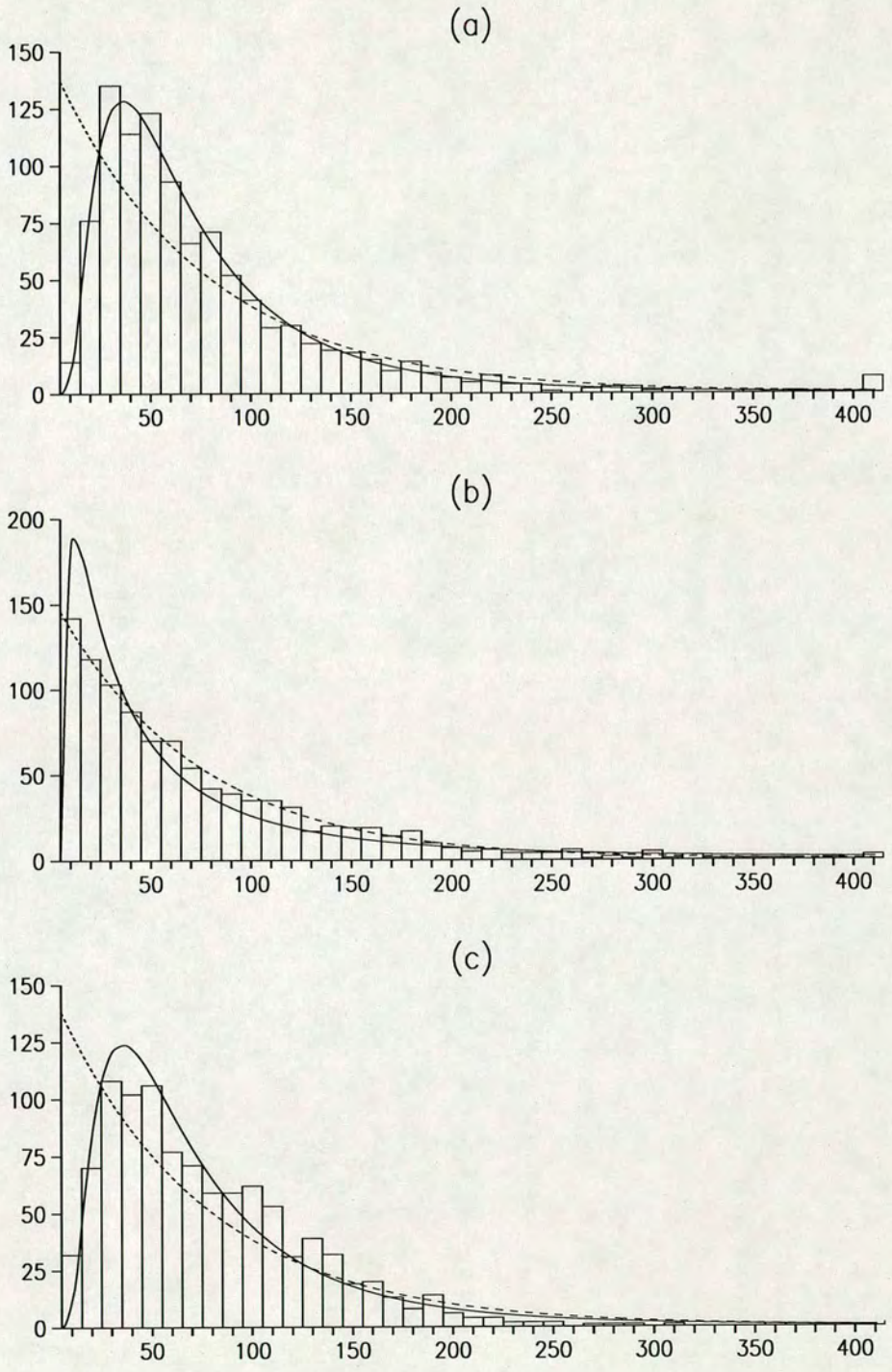

Figure 7.1: *Histograms of simulated data under (a) log-normal model, (b) exponential model, (c) gamma model; (—) fit of log-normal model, (- - -) fit of exponential model.*

- Log-normal data: no evidence of departure from log-normal model; strong evidence of departure from exponential model in the direction of the log-normal.
- Exponential data: strong evidence of departure from log-normal model in the direction of the exponential; no evidence of departure from the exponential model.
- Gamma data: strong evidence of departure from the log-normal model in the direction of the exponential; strong evidence of departure from the exponential in the direction of the log-normal.

In the first two cases we have the expected result. In the third case it would appear that the correct model is somewhere 'in between' the exponential and log-normal models, which from the histogram is seen to be an entirely plausible conclusion, although we can note that the log-normal appears a better approximation than the exponential, which is not reflected in the relative sizes of T_1^* and T_2^*.

7.2.1.2 Other related approaches

Atkinson (1970) investigated the exponential combination of the two hypotheses, producing a model which embeds both competing models as special cases, i.e.

$$f_1(x, \alpha)^{\lambda} f_2(x, \beta)^{(1-\lambda)}.$$

He then considered evidence for whether $\lambda = 0$ or $\lambda = 1$. He also considered the hypothesis $\lambda = 1/2$, as a test of equidistance between the two models, but the interpretation of the value $1/2$ was questioned in the subsequent discussion, as it is not clear that there is any reason to see this value as being the special point that is halfway between the two models. Atkinson also derived a statistic asymptotically equivalent to Cox's, viewable as a Lagrange multiplier or score test statistic, which replaces $\hat{\beta}$ by $\beta_{\hat{\alpha}}$ in T_1. However a simulation study was unable to conclude which is preferable, both approaching asymptotic normality disappointingly slowly. White (1982) provided a third version of the statistic, replacing $\beta_{\hat{\alpha}}$ in the second term of T_1 by $\hat{\beta}$. In all these cases the corresponding replacements are made in T_2. Victoria-Feser (1997) discussed the three versions and proposed a robust version, which she called a generalised Lagrange multiplier test, the computation of which is far from straightforward. Kent (1986) proposed yet another statistic and attepted to clarify the relationships between the tests geometrically.

7.2.2 Bayesian approach

Similar ideas of model comparison have been considered within a Bayesian framework. Gelman et al. (1995, Chapter 6) discussed model checking by simulating values from the posterior predictive distribution and comparing with the observed data, systematic differences indicating failure of the model. Instead of test statistics $T(x)$, dependent only on the data, test quantities $T(x, \theta)$ can be considered, generalising test statistics to allow dependence on the posterior distributions of the parameters also. For comparison of a set of discrete (separate) models, the Bayes factor $B(x)$ may be useful, i.e. the ratio of the marginal likelihoods under the two models, which quantifies the evidence in favour of one model over the other. If our data are x and we are comparing Models H_1 and H_2 then

$$\frac{p(H_2|x)}{p(H_1|x)} = \frac{p(H_2)}{p(H_1)} \times \frac{l(x|H_2)}{l(x|H_1)}$$

i.e. ratio of posterior probabilities = ratio of prior probabilities $\times$ Bayes factor. We can also write

$$l(x|H_j) = \int l(\theta_j) p_j(\theta_j) d\theta_j$$

for $j = 1, 2$, hence the Bayes factor can also be thought of as the ratio of 'prior means' of likelihoods.

Several variants on this theme have been proposed. Aitkin (1991) commented that the Bayes factor is very sensitive to variation in the priors and defined the posterior Bayes factor as the ratio of posterior means, i.e. as a ratio of terms of the form

$$l^*(x|H_j) = \int l(\theta_j) p_j(\theta_j|x) d\theta_j,$$

which is claimed to reduce sensitivity to variations in the prior. However Kass and Raftery (1995) claimed that the procedure has little Bayesian justification and can lead to counterintuitive results. O'Hagan (1995) went on to propose the fractional Bayes factor, a variant of the partial Bayes factor. Partial Bayes factors are so-called as they are based on only part of the data. The full data x are divided into two parts, the training sample y and the part used for the comparison z. The partial Bayes factor is given by

$$B(z|y) = \frac{l_1(z|y)}{l_2(z|y)}$$

where the priors must be proper. Here, $l_j(\cdot)$ is the likelihood given model H_j. The full Bayes factor is given by

$$B(x) = B(y)B(z|y).$$

Different training samples can be selected and results averaged. However if there are n observations altogether and m in the training sample then all combinations should be used and Bayes factors averaged. It is the lack of any obvious way to average that motivates the fractional Bayes factor $B_b(x)$, given by

$$B_b(x) = \frac{l_1(b, x)}{l_2(b, x)}$$

where $b = m/n$. This is asymptotically (large m and n) equivalent to the partial Bayes factor but is also proposed as an alternative even for m and n small. A generalisation of this provides a version which has the full, fractional and posterior Bayes factors all as special cases.

O'Hagan (1995) applied the fractional Bayes factor to non-nested models, including the log-normal vs exponential example that we consider in the next section, and demonstrated the four possibilities that cumulative $\log B_b(x)$ can display, i.e.

- tend to ∞, indicating Model 1 is 'more correct',
- tend to $-\infty$, indicating Model 2 is 'more correct',
- behave erratically, indicating neither model is correct,
- stay close/converge to 0, indicating both models are equally good.

Finally it should be noted that Schwarz's Bayesian information criterion (BIC) is actually an asymptotic approximation for the Bayes factor.

7.2.3 Simulation/parametric bootstrap approaches

In simple cases, the forms of the Cox statistics T_1 and T_2 can be derived explicitly, e.g. for the comparison of the fit of two distributions from the exponential family to a random sample of observations (Cox, 1961). However for more complicated situations the evaluation of the statistic can be prohibitive. Simulation therefore can be utilised, and a bootstrap-style p-value will avoid the problem of the Cox statistics being slow to approach asymptotic normality.

Simulation approaches were first considered in the 1950s, but are now a far more realistic alternative with the increased computing power available. We therefore consider techniques similar to those employed by Williams (1970), Hinde (1992) and Ross (1998). In comparing two competing models, the basic idea is to fit each model according to some fitting criteria, e.g. maximum likelihood or least squares. We then simulate from each of the models with parameters as estimated, and re-fit both models to both sets of simulated data. By comparing the values of

fitting criteria for each set of simulated data under both models with the values for the observed data, we can see which model the data are more consistent with. Williams (1970) used this approach, considering residual sums of squares for two regression models, Atkinson (1985) termed it a Monte-Carlo test and applied it to the choice between two probability distributions, Hinde (1992) applied it to generalised linear models and Ross (1998) considered non-linear regression models. The method can also be termed a parametric bootstrap as described in Efron and Tibshirani (1993, Chapters 6 and 21) and Davison and Hinkley (1997, Section 4.2).

All the above consider comparison of two models. If there are three or more competing models it would be useful to be able to compare them all simultaneously, rather than pairwise. This can be done by extension of the methodology described above. The steps are as follows.

1. Estimate parameters for Models A, B and C using observed data.
2. Simulate realisations (e.g. 100) for each model with parameters as estimated.
3. For each set of simulated data, re-fit all three models.
4. Compare the relative sizes of the fitting criteria for each set of simulated data under all three models with those of the observed data.

Variations on this would be for Step 2 to consider the Bayesian approach and instead of simulating from the models with fixed point estimates (e.g. maximum likelihood estimates), to use parameters from the posterior distributions. For Step 3 an alternative, considered below, is instead of re-fitting all the models to the simulated datasets, simply to evaluate the likelihood (or other fitting criterion) under the parameters as estimated from the data. We show below that this can be a good approximate alternative, although lacking the theoretical basis of re-maximising the likelihood.

7.2.3.1 Return to example from Cox (1961)

We illustrate the parametric bootstrap approach by considering the example from Cox (1961), as considered in Section 7.2.1.1. There, we used the Cox statistics to test whether an observed sample came from an exponential or a log-normal distribution. We now use the same example to illustrate the parametric bootstrap approach. Figure 7.2 illustrates the results obtained. Plots labelled (i) show the values of the maximised likelihood for the log-normal model against the exponential model and plots labelled (ii) show the likelihood evaluated under the

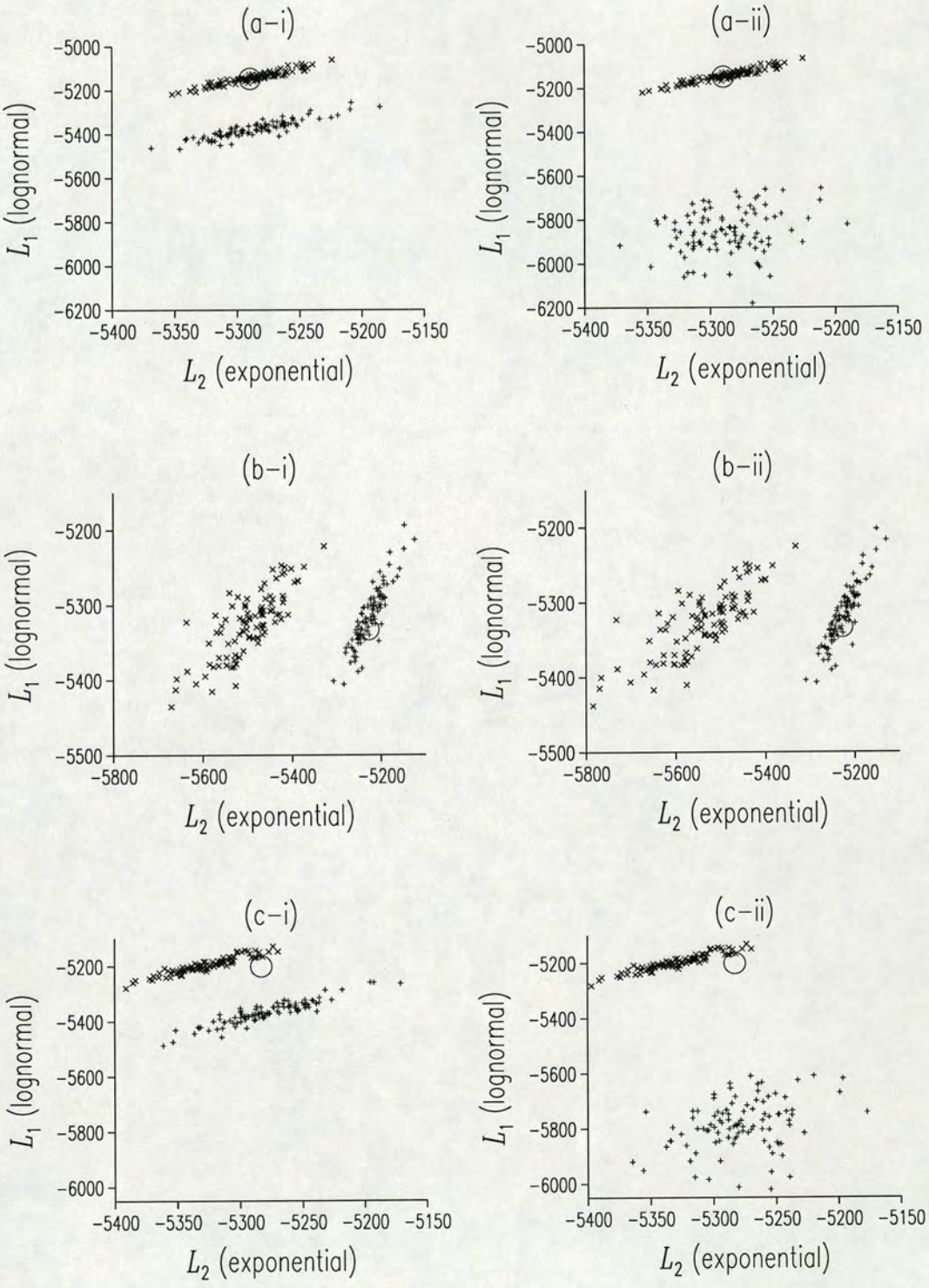

Figure 7.2: *Log-likelihoods under both log-normal and exponential models; (×) data simulated from the fitted log-normal model, (+) data simulated from the fitted exponential model, (◯) 'observed' data; (a) data are log-normal, (b) data are exponential, (c) data are gamma; (i) maximised likelihoods, (ii) likelihoods evaluated under fitted parameters.*

maximum likelihood estimates from the observed data. Both sets of plots are as we would expect. For (a) the data fall within the range covered by the simulated data from the log-normal distribution, and for (b) the data fall within the range covered by the exponential distribution simulations. For (c) the data fall outwith both regions, although much closer to the log-normal simulations, reflecting what we saw in the histogram of Figure 7.1(c). Note that this closer fit to the log-normal was not reflected in the Cox statistics, which were of a similar size for departures from both log-normal and exponential distributions. It should be borne in mind however that the Euclidean distance on the plots is not necessarily meaningful, and so although the data look closer to one set of simulations than the other, this does not necessarily imply a closer fit to one model than the other. We should really produce a p-value for the consistency with each set of simulated data. In the next section we discuss the potential use of bootstrap p-values and order statistics in more than one dimension.

It is useful also to compare the relative pictures given by (i) and (ii) in Figure 7.2. The same conclusions are drawn from consideration of either set of figures. The only difference is that the simulated data for the less well-fitting model generally have a greater spread when the likelihood is evaluated under the data-estimated parameters (ii) rather than the maximum likelihood for each simulated series (i). Option (i) is better rooted in the theory, but for more complicated models where there might be issues of either computational slowness to re-maximise the likelihood or difficulty of obtaining reliable estimates, it appears that the likelihood simply evaluated under the data-estimated parameters will give a good approximation to the situation. We will return to this comparison for the cow feeding data later on.

7.2.3.2 More than two competing models

For two competing models, a two-dimensional plot can be used to assess whether the data fall within the expected region with regards to the two likelihood values. However, using Figure 7.2(c-i) as an example, it can be appreciated that inspection of the two marginal likelihoods would not be sufficient, as here this would indicate that the data lie within both marginals for the log-normal simulated data, whereas in the bivariate view the data lie outwith the simulated data. Therefore for comparison of three competing models, inspection of the three projections onto each pair of axes might indicate the data being consistent with simulations, but it is easy to envisage the spread of the simulated data taking the shape of a disc angled diagonally, with the data falling outside, but the three projections

onto the axes would show the data to be consistent with the simulations.

Therefore for comparison of a set of r models, we have to consider the likelihoods in r-space and not simply the projections onto pairs of axes. One approach would be to construct a convex hull around the set of points in r-space and strip these down in order to obtain a p-value. A discussion of this, and other ideas of rectangular peeling and elliptical peeling, is given in Green (1981). For elliptical shapes the Mahalanobis distance could be used as a measure. It is also thought that projection along principal components or use of other multivariate techniques may give a more useful summary. These ideas remain to be investigated although we illustrate use of principal components below.

7.3 Results for cow feeding data

For each of the eight high-protein cows we have considered the models described in Section 7.1. We now want to apply the simulation methods described to see which types of model fit the data best according to these criteria. Although in theory we could consider the comparison of all models simultaneously, we shall consider the three classes — latent Gaussian, hidden Markov and semi-Markov — separately, and then, having selected the most appropriate model from each class, compare the resulting three models. We also compare use of the likelihood evaluated at the maximum likelihood estimates given by the data, with re-maximisation of the likelihood for each simulated series. We consider results for Cow 5 in detail as an example and then summarise results for the rest of the high-protein cows.

7.3.1 Hidden Markov models

We consider the two- and three-state hidden Markov models and discrete-time compartment models. Parameter estimates are given in Tables 5.1, 5.4 and 5.6, and values of AIC and BIC in Tables 5.5 and 5.7, with a summary for the three-state models in Table 5.8. For Cow 5 we concluded that the discrete-time compartment model was the most appropriate of the models considered. We now check that the methods developed here give the same conclusions, by comparing the two- and three-state hidden Markov models and the three-state discrete-time compartment model. Figure 7.3 shows comparisons for when we simply evaluate the likelihoods for the simulated series at the maximum likelihood estimates obtained from the data, and Figure 7.4 shows the corresponding picture when likelihoods are re-maximised for all the simulated series. The figures give almost iden-

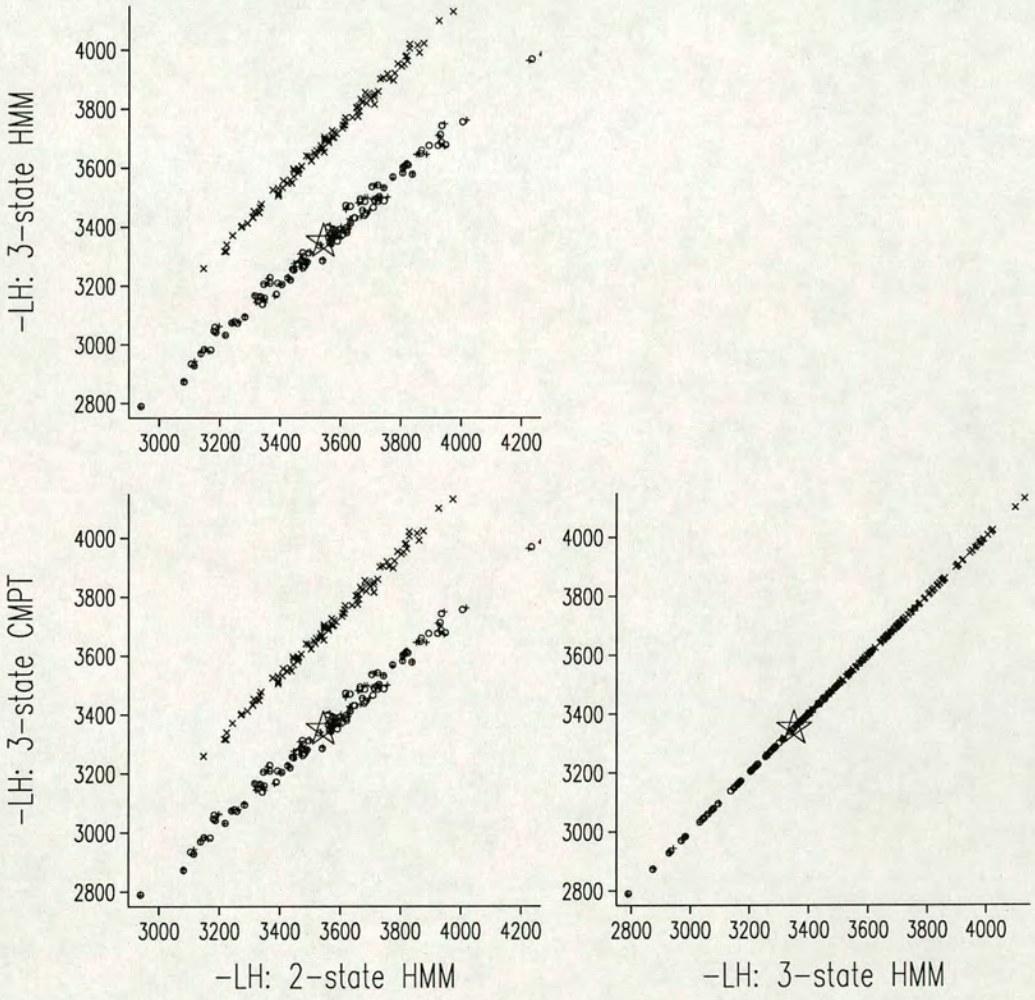

Figure 7.3: *Cow 5; negative log-likelihoods evaluated at the maximum likelihood estimates of parameters from the observed data, for series simulated under (×) two-state HMM, (○) three-state HMM, (+) three-state compartment model; (⋆) data.*

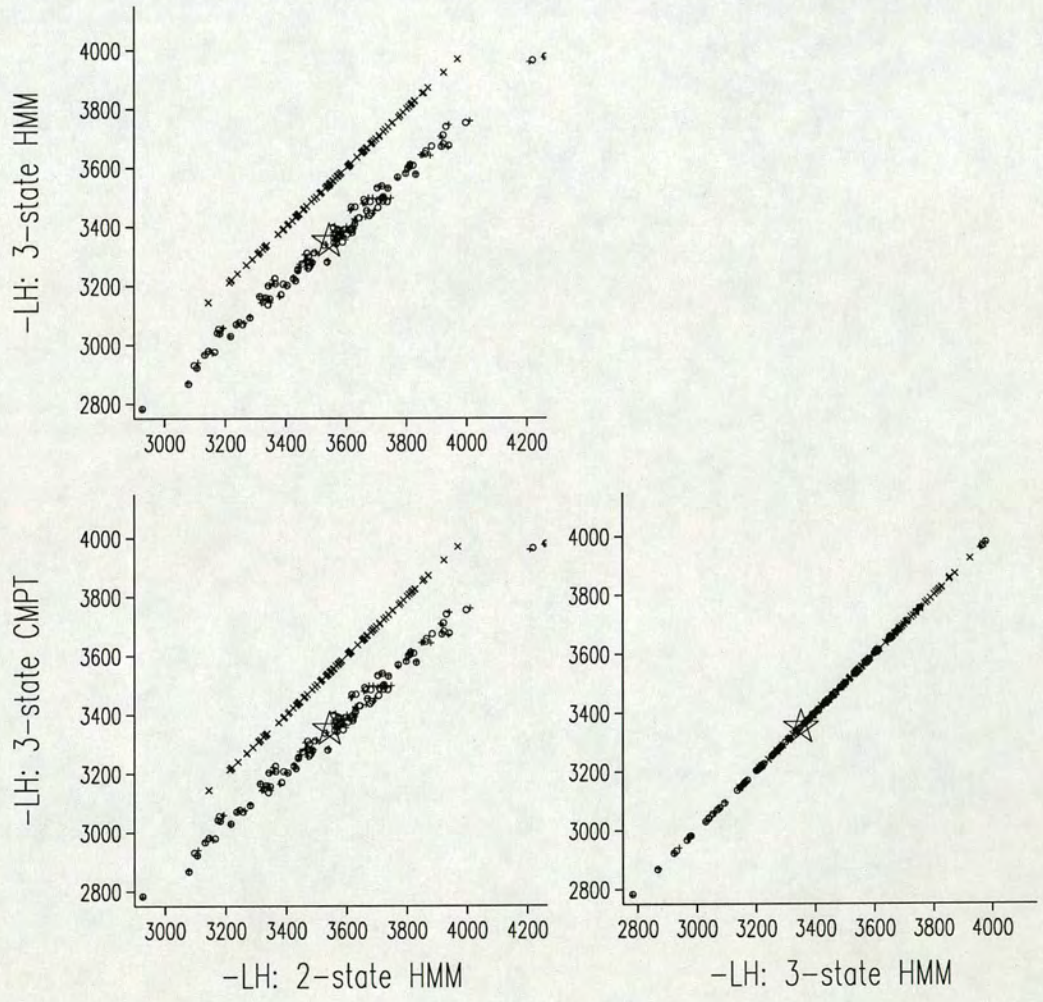

Figure 7.4: *Cow 5; negative log-likelihoods for models re-fit to series simulated under (×) two-state HMM, (○) three-state HMM, (+) three-state compartment model; (⋆) data.*

tical pictures and show the data to be consistent with both the three-state hidden Markov model and the three-state discrete-time compartment model. Therefore in the interests of model parsimony the compartment model is preferred. This is consistent with the results of Tables 5.5 and 5.7 which show the likelihood to be the same under both of the three-state models, but the information criteria to be lower for the compartment model due to the smaller number of parameters to be estimated. For the other cows, the three-state compartment model was found to be adequate for Cows 41 and 169, whereas the three-state hidden Markov model seemed better for the remaining five cows (108, 170, 182, 194 and 221). These results can be reconciled with the figures in Tables 5.5 and 5.7, though sometimes even when the value of BIC was slightly lower for the three-state hidden Markov model, the simulation method showed the compartment model to be adequate.

7.3.2 Semi-Markov models

Here we compare the semi-Markov models for which parameter values are given in Tables 6.1 and 6.2 and values of likelihood, AIC and BIC in Table 6.3. From these we see that for Cow 5 we might expect the three-state model to be adequate, although BIC was lower for the four-state model. Figure 7.5 shows the simulation method applied to these methods, showing both models to be similar and the three-state model to be adequate. The likelihoods plotted in this figure are those evaluated under the parameters estimated from the data. If we try and re-maximise likelihoods there are problems with identifiability when trying to fit the four-state model to data simulated from the three-state model. To overcome this I have fixed the three means and variances of the distributions and just re-estimated the transition probabilities. These results are shown in Figure 7.6. Although the identifiability problems mean that the likelihood has not been completely re-maximised, we can still note the strong similarity with Figure 7.5. Hence from either graph it would have to be concluded that both models are acceptable and therefore we choose the simpler three-state model as the preferred one, agreeing with previous findings. For the other cows, similar procedures showed a three-state model to be adequate also for Cows 41, 169 and 170, whereas for the rest (108, 182, 194 and 221) the four-state model offers some benefit. These results can be reconciled with Table 6.3, though again sometimes we see slightly lower values of BIC for the four-state model, even though the simulation shows the three-state model to be adequate.

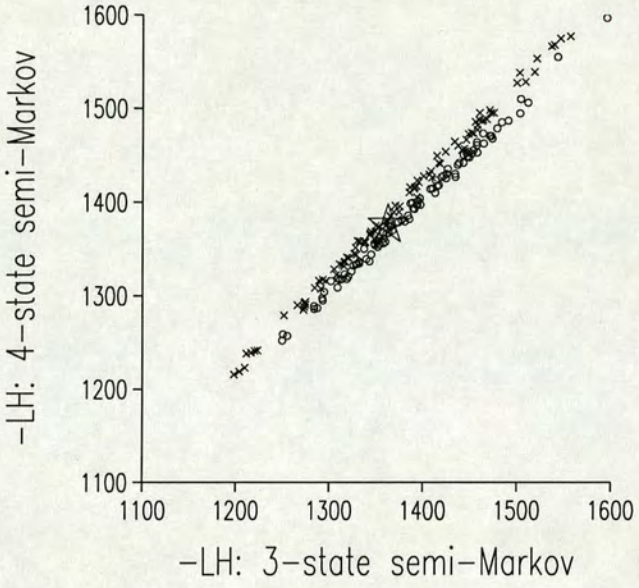

Figure 7.5: *Cow 5; negative log-likelihoods evaluated at the maximum likelihood estimates of parameters from the observed data, for series simulated under (×) three-state semi-Markov model, (○) four-state semi-Markov model; (⋆) data.*

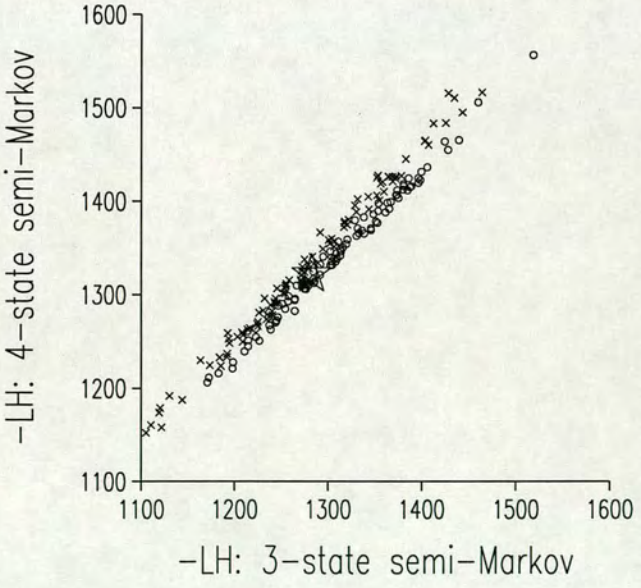

Figure 7.6: *Cow 5; negative log-likelihoods for models re-fit to series simulated under (×) three-state semi-Markov model, (○) four-state semi-Markov model; (⋆) data.*

7.3.3 Comparison between hidden Markov, semi-Markov and latent Gaussian models

We now compare the most appropriate hidden Markov and semi-Markov models, both with each other and with the latent Gaussian model. For Cow 5 the three-state discrete-time compartment was chosen as the preferred model from the class of hidden Markov models and the semi-Markov with three states was seen to be adequate. Figure 7.7 therefore plots the values of the fitting criteria evaluated at the estimates given by the data, and Figure 7.8 shows the same plots when the models are re-fitted. Note that these fitting criteria are log-likelihoods for the hidden Markov and semi-Markov models, but are sums of squares for the latent Gaussian model. We have chosen to display the sum of squares using the binary autocorrelation to fit the latent Gaussian model, but use of the Gaussian autocorrelation produces very similar pictures. These figures show firstly that the data are inconsistent with the latent Gaussian model. It shows the hidden Markov and semi-Markov models as being much more similar to each other and hence the regions of the graph containing these simulated data are overlapping. Nevertheless, the top plot of Figure 7.7 shows the data to be consistent with the semi-Markov model and not with either of the others. We have already noted that we should really be looking at a three-dimensional plot of the arrangement of points rather than these three projections. In Figures 7.9 and 7.10 we present the results in terms of principal components, calculated using the correlation matrix, as the fitting criteria are on different scales. Therefore Figure 7.9 shows similar information as Figure 7.7, with scales standardised and from different angles, and similarly Figures 7.10 and 7.8. For both we can see that the data are more consistent with the semi-Markov model than with the other models, but perhaps less convincingly in Figure 7.10.

Considering a different cow, Cow 182, we found the three-state hidden Markov model was preferable to the other hidden Markov/compartment models, and the four-state semi-Markov model was preferable to that with three states. Therefore we compare these two models with each other and with the latent Gaussian model in Figures 7.11 and 7.12, again showing evaluated and re-maximised likelihoods, respectively. All plots here confirm that again the observed data are consistent with the semi-Markov model, but not with the other two models.

Obviously this does not confirm that the semi-Markov model is a perfect fit to the data, but from the criteria examined, we see no inconsistencies with what would be expected if the observed data did arise from this model. Many other

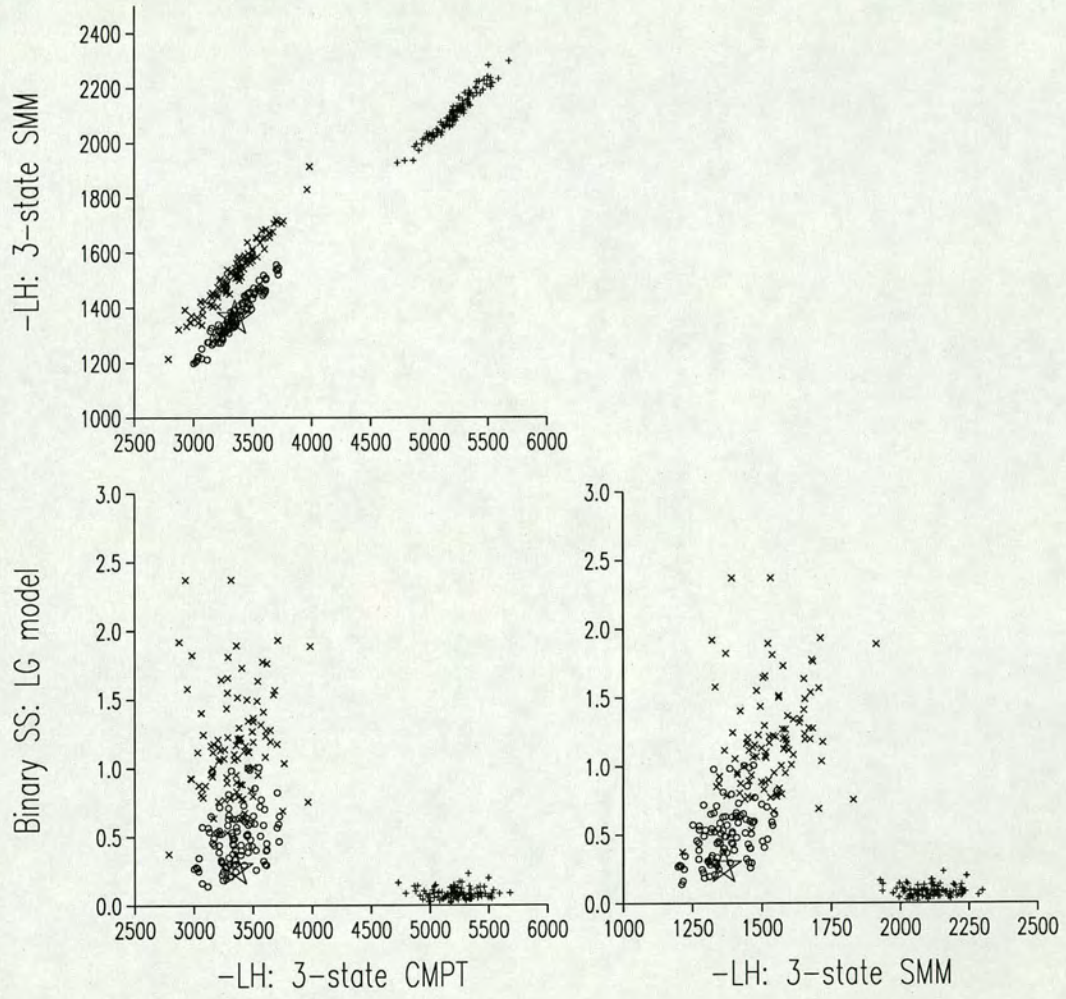

Figure 7.7: *Cow 5; fitting criteria evaluated at the parameters estimated from the observed data, for series simulated under (×) three-state discrete-time compartment model, (○) three-state semi-Markov model, (+) latent Gaussian model; (⋆) data.*

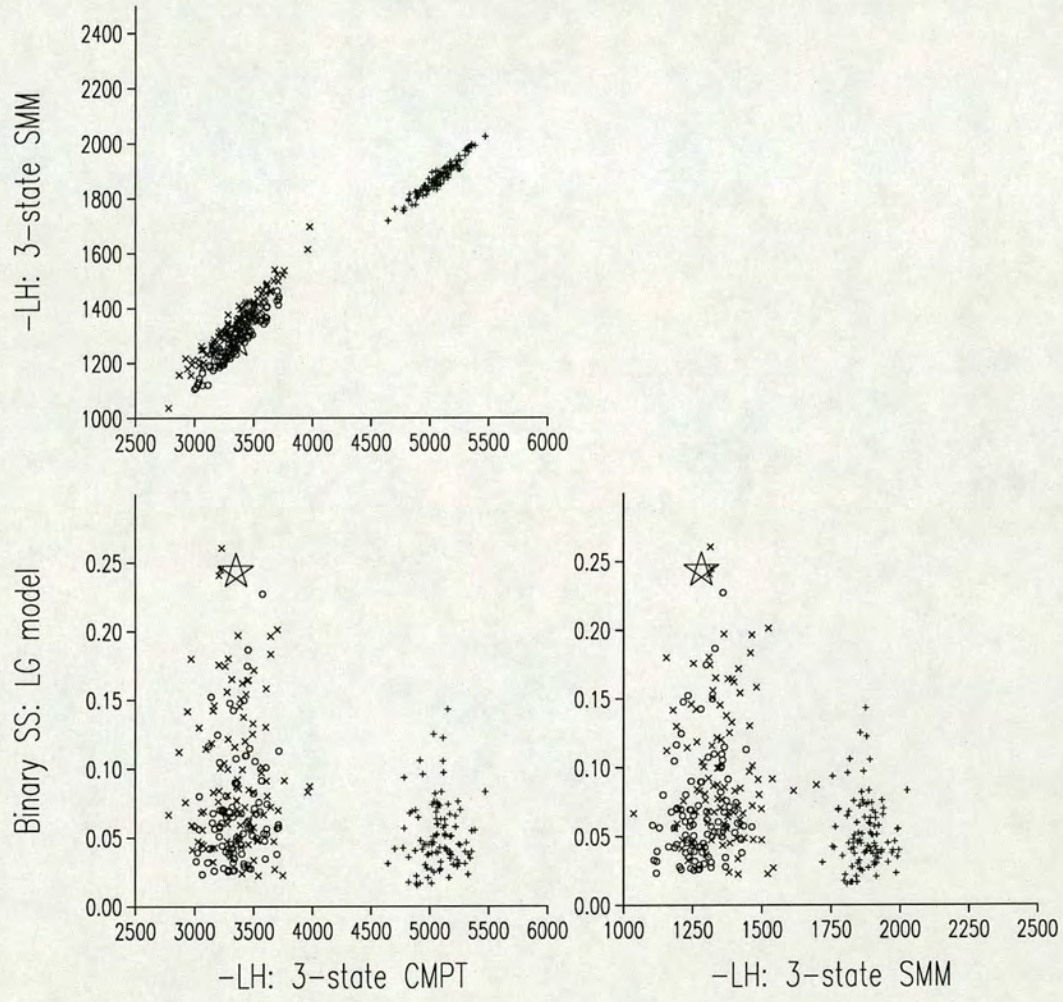

Figure 7.8: *Cow 5; fitting criteria for models re-fit to series simulated under (×) three-state discrete-time compartment model, (○) three-state semi-Markov model, (+) latent Gaussian model; (⋆) data.*

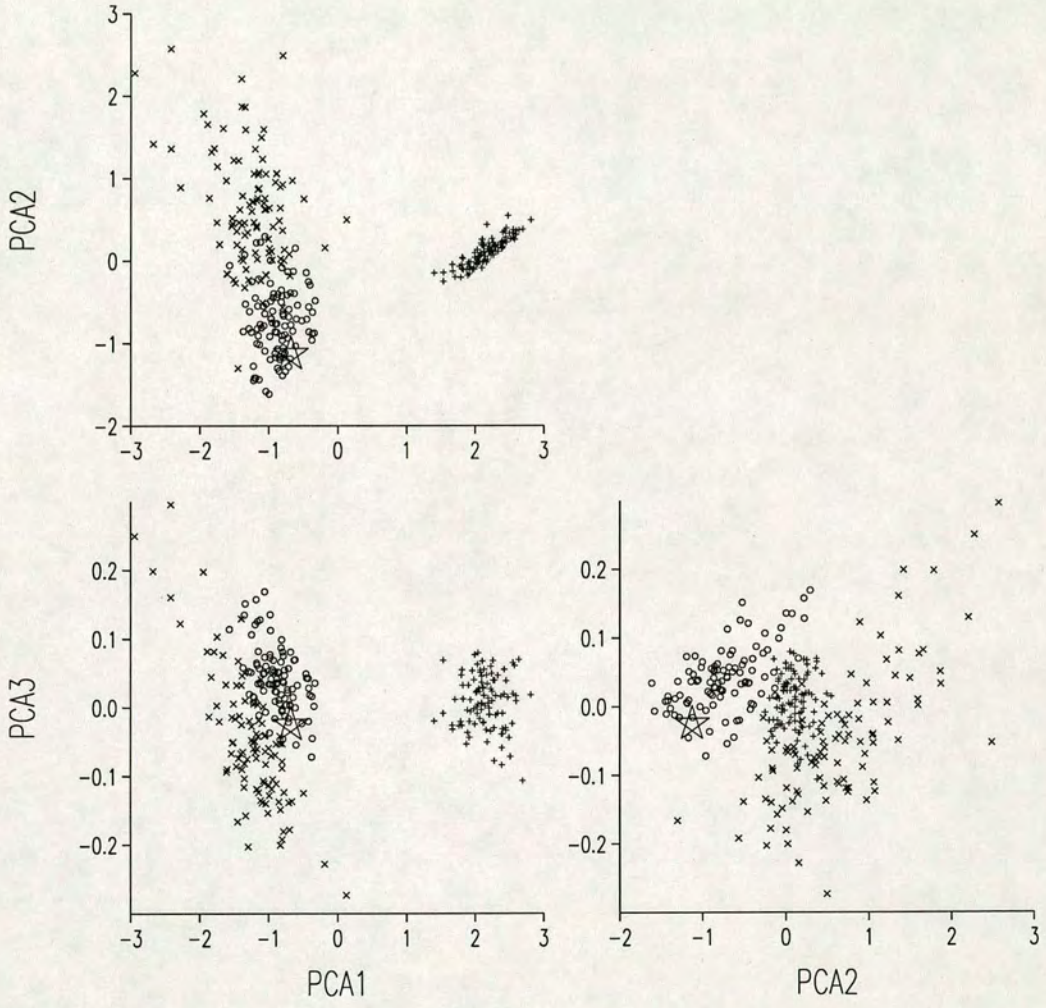

Figure 7.9: *Cow 5; principal components of the fitting criteria evaluated at the parameters estimated from the observed data, for series simulated under (×) three-state discrete-time compartment model, (○) three-state semi-Markov model, (+) latent Gaussian model; (⋆) data.*

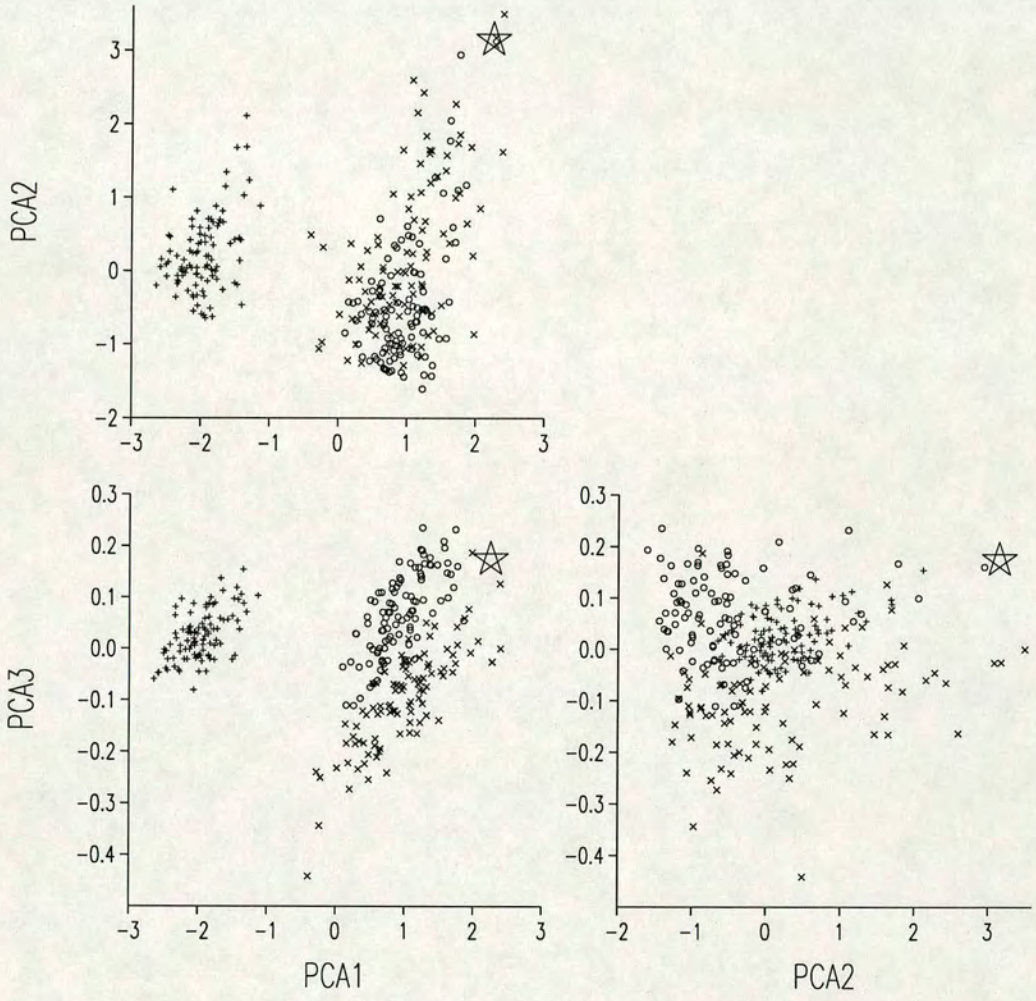

Figure 7.10: *Cow 5; principal components of the fitting criteria for models re-fit to series simulated under (×) three-state discrete-time compartment model, (○) three-state semi-Markov model, (+) latent Gaussian model; (⋆) data.*

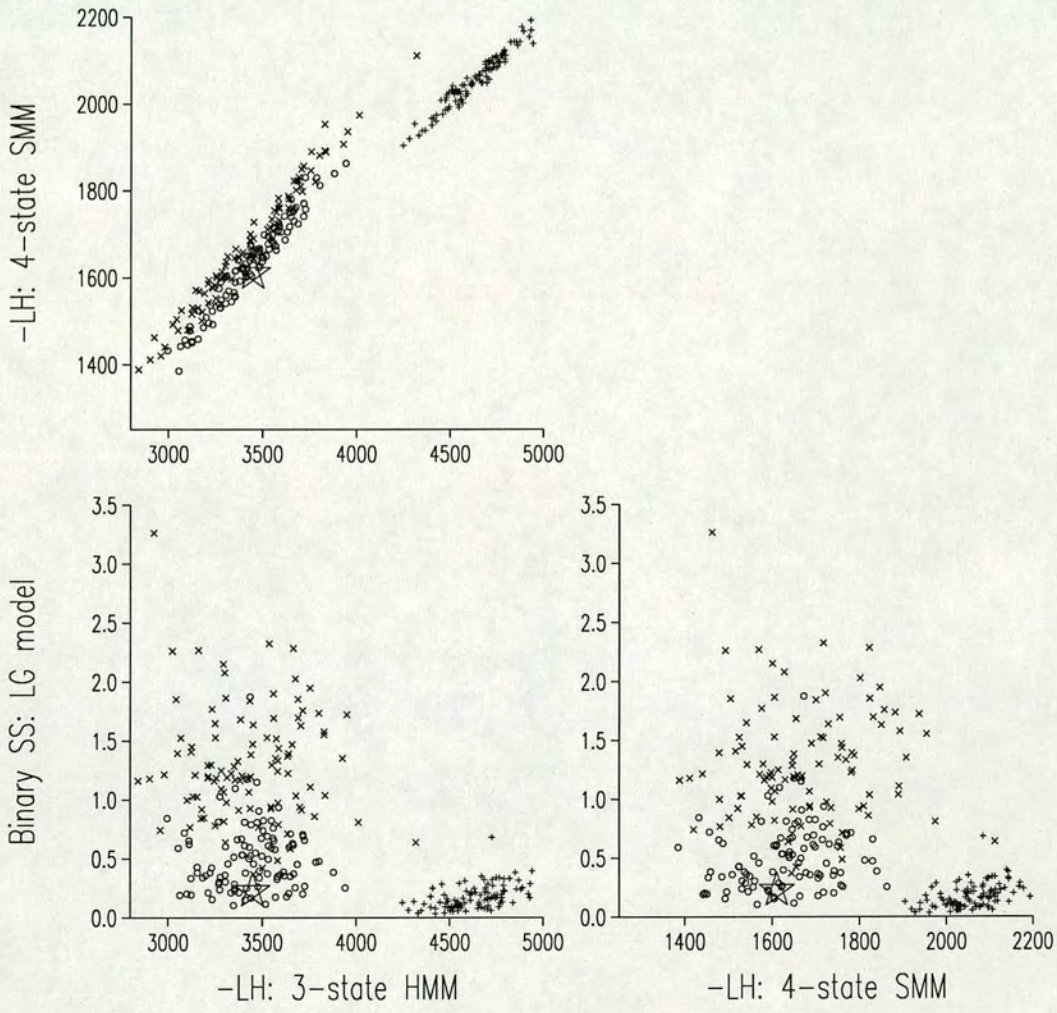

Figure 7.11: *Cow 182; fitting criteria evaluated at the parameters estimated from the observed data, for series simulated under (×) three-state hidden Markov model, (○) four-state semi-Markov model, (+) latent Gaussian model; (⋆) data.*

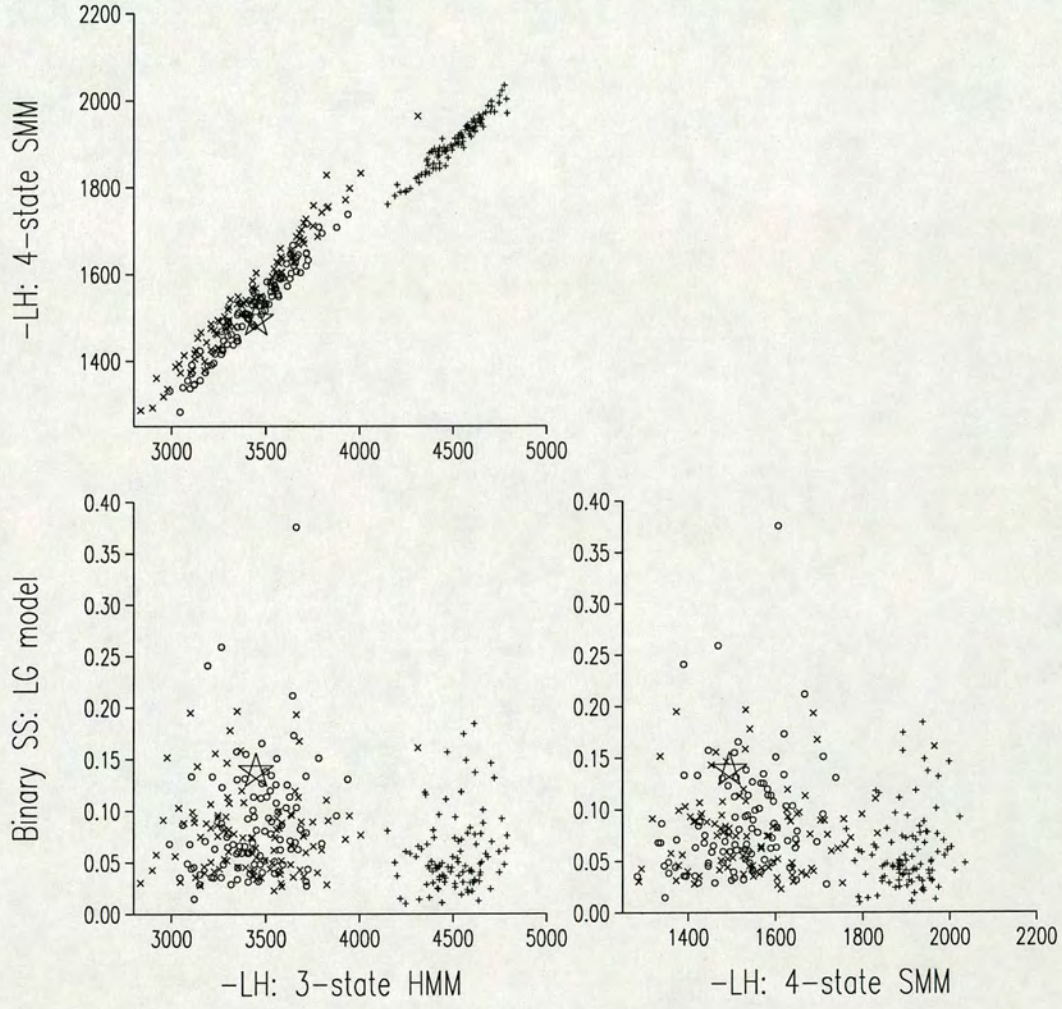

Figure 7.12: *Cow 182; fitting criteria for models re-fit to series simulated under (×) three-state hidden Markov model, (○) four-state semi-Markov model, (+) latent Gaussian model; (⋆) data.*

Cow	*Hidden Markov*			*Semi-Markov*		
	Preferred	D_A	D_B	*Preferred*	D_A	D_B
5	3-state cmpt	−6	−33	3-state	18	−8
41	3-state cmpt	0	−26	3-state	22	−2
108	3-state HMM	48	25	4-state	113	83
169	3-state cmpt	−5	−31	3-state	23	−3
170	3-state HMM	20	−6	3-state	44	15
182	3-state HMM	10	−16	4-state	44	17
194	3-state HMM	5	−21	4-state	44	18
221	3-state HMM	52	26	4-state	55	22

Table 7.3: *Preferred models of hidden Markov and semi-Markov types for the eight high-protein cows. D_A and D_B are the differences between AIC and BIC, respectively, either for the three-state hidden Markov vs compartment models, or for the three-state vs two-state semi-Markov models. In either case a positive figure indicates better fitting of the more general model.*

quantities could also be used as criteria to check for consistency, but we argue that the actual criteria used to fit the model is the natural choice.

7.3.4 Summary of results for high-protein cows

Here we give a short summary for all the high-protein cows. Table 7.3 shows which of the hidden Markov and semi-Markov models are the preferred model for each of the cows. For the hidden Markov models, the differences in AIC and BIC shown refer to the three-state HMM compared with the three-state compartment model, a positive value indicating the criterion taking a lower value for the hidden Markov model and hence being the preferred one. For semi-Markov models we just compare the three- and four-state models and here a positive value for the differences indicates that the four-state model is the better choice. With regards to the hidden Markov models, for Cows 5, 41 and 169, the three-state discrete time compartment model is the best of those considered, using either of the criteria. For the other five cows there is some evidence that the three-state HMM offers a better description. For the semi-Markov models, three states appear to be adequate for Cows 5, 41 and 169, whereas there is evidence that the four-state model offers some benefit for Cows 108, 170, 182, 194 and 221. Note that these conclusions are based not just on the AIC and BIC, but also on the simulation methods of this chapter, even though graphs are not shown for every cow. In the case of the semi-Markov models, plots of the marginal distributions of non-feeding durations are also useful in deciding whether two or three non-feeding states are required.

More work needs to be done in order to base conclusions on a thorough inspection of the graphs from all aspects in three dimensions, and the ideas described in Section 7.2.3.2 of data peeling could be used to gain a p-value for how consistent the data are with each model. However this is left as further work and here we assume that the plots presented are adequate on which to base conclusions.

7.4 Summary

I have discussed the connections and differences between the models developed in earlier chapters, in terms of whether the models occur in a continuous or discrete time framework, their latent structure, time-dependency and ability to incorporate diurnal patterns. Some of the methods from the literature on comparison of non-nested models were then reviewed, and methods to compare models developed. Simulation-based methods were shown to be capable of comparing the fit of unrelated models, even when fit in different ways and according to different fitting criteria. The approach involves fitting the different models, simulating many realisations of data from each fitted model and then re-fitting all models to each set of simulated data. It can then be evaluated how consistent the observed data are with the simulations from each model. I also showed that instead of re-fitting all models to each set of simulated data, simple evaluation of the likelihood under the parameters estimated from the data gives the same picture in the cases considered. Applying these methods to the cow feeding data showed that, of the models considered, semi-Markov models appear to be most consistent with the data. Further work needs to be done to develop formal tests of which models fit best, and ideas relating to order statistics in higher dimensions could be employed to allow more formal comparison of several models simultaneously.

Chapter 8

Discussion and further work

In this final short chapter, I review some main points in relation to the objectives set out in Chapter 1, before going on to suggest areas for further work.

8.1 Review of objectives

In this thesis I have attempted to identify some useful approaches to modelling animal behaviour data. All the work has been based around a large set of cow feeding data, but many of the ideas would be equally relevant to other types of behaviour data, and to other types of binary and categorical data. I firstly summarise the overall structure of the project.

- I looked at past work on the modelling of animal behaviour data and, together with results from initial exploration of the cow feeding data, identified some modelling strategies to progress. In order to be potentially useful, models needed to be plausible both from a biological viewpoint and in terms of any necessary statistical assumptions.
- Methodology was developed for the fitting of the models to the data, investigating the statistical properties and assessing whether these were suitable for the given data. The fit of the observed data to the model was illustrated and comparisons made with realisations from the fitted models. Even if there were drawbacks for the fit of models to these particular datasets, they are still potentially useful for other datasets.
- The fit of the models was compared and issues addressed of why certain models should fit better than others. The statistical properties of the best-fitting models then need to be related to the underlying biological processes that produced the observed behaviour in the first place.

The third point here has not been fully addressed, but it can be argued that at this stage the work requires more collaborative thinking with biologists.

From a statistical point of view, novel methodology has been developed and models have been fit that have not previously been considered for animal behaviour data.

- For the latent Gaussian model I demonstrated how ARMA models may be fit to censored data, in particular binary data, by considering the autocorrelation structure of the observed binary series and noting the one-to-one correspondence with the autocorrelation of the Gaussian series assumed to be underlying. The likelihood for a Gaussian process was expressed in its spectral form, and I proved that the full likelihood may be approximated by using far fewer terms than the full set. This estimator was considered alongside other computationally fast estimators in a simulation study which compared them with each other and in relation to an MCMC method for which I also developed methodology.

- For hidden Markov models, no new statistical methodology was developed, but I am aware of only one published paper which uses these models for animal behaviour, MacDonald and Raubenheimer (1995), who fit the models to behaviour of locusts. Hence it was interesting to investigate how much potential these models have for describing the cow feeding dataset.

- The semi-Markov models fit were non-standard in that the states were not fully observed; more than one non-feeding state was being assumed, and hence when the cows were in a non-feeding period, the current state was unknown. I showed how the EM algorithm could be used to fit models to these type of data. In terms of the algorithms for estimating parameters, the methodology parallels that used to fit hidden Markov models. The method is far more satisfactory than a multi-stage procedure for which events would have to be first classified into states and then models fit which would fail to take into account the uncertainty involved in the initial classification.

- The comparison of separate (non-nested) models is still a problem for which many statisticians are unclear as to the best procedure. Therefore I have reviewed the literature in this area and advocated a simulation-based approach, also termed parametric bootstrapping. This not only allows the comparison of models that are non-nested, but can cope with models that are fit according to different criteria. Hence it was possible to compare models fit by least squares with those fit by maximum likelihood, and even

models fit to different forms of the same dataset. Here, the semi-Markov model was fit to the original dataset in the form of a sequence of behaviours and their durations, whereas the other models were fit to the data after being discretised. Hence it was possible to use this method to compare models formulated either in discrete or continuous time.

It was seen that semi-Markov models came out overall as being statistically the best fit to the data. This was probably due to this type of model being the one that most accurately captured the marginal distributions of the behavioural events. Indeed, along with the form of its dependence structure, the direct specification of the marginal distributions was the main motivation for this class of models. The latent Gaussian model was seen to produce marginal distributions that were at least of the correct shape, whereas the hidden Markov model was restricted to having marginal distributions that were mixtures of geometric distributions which, in this case at least, was not adequate.

In terms of their fundamental structure, I still think the latent Gaussian model has a good plausible biological motivation. This is the only model out of those considered to have a continuous latent variable, and this is still an attractive feature biologically. When a cow passes from non-feeding to feeding, this must happen as a result of some increasing imbalance within the cow. Once a certain time has passed since the last meal, she will generally start to feel hungry, and this hunger will increase until it has reached a sufficient level for her to do something about it and begin feeding. It would seem natural to model this hunger tendency with a continuous variable, rather than as a discrete variable which instantaneously switches from her being not hungry to being hungry. Therefore surely a model which is trying to explain the underlying mechanisms of the feeding would have to incorporate this hunger tendency. It is a strong assumption that this underlying variable is of a stationary Gaussian form, but from a statistical point of view this is a convenient place to start.

For the semi-Markov model, the idea of hunger tendency corresponds to the hazard function for the marginal distribution of the current event type. By modelling non-feeding periods with log-normal distributions, the hazard increases to a maximum before decreasing again. The interpretation of the decrease was that, given an animal has not eaten for several hours, the probability of her eating in any given minute starts to decrease. If this is not deemed a sound biological feature, use of other distributions, e.g. Weibull, can produce an always-increasing hazard function. For any Markov model, as opposed to semi-Markov, hazard functions are always constant and therefore tendency to perform a particular behaviour

next does not depend on how long the current behaviour has lasted for. This is a fault of the hidden Markov model, for which the main attraction was that it is the underlying state of the animal that is being directly modelled, rather than the observed behaviours. However it has also been noted that although the semi-Markov model does not do this directly, after fitting the model, we can aggregate the within-meal states into a 'meal' state, producing a latent-like structure not dissimilar to the hidden Markov model.

The option of modelling meal data rather than visit data was discussed, and although some authors had decided that meal data were preferable in that it might more accurately describe the overall behaviour and be less prone to herd and dominance effects, we chose to directly model the individual feeder-visit data. This was mainly so that intra-meal patterns could also be modelled and, as discussed above, to model the meal data requires an initial classification of inter-feeding durations as within- and between-meal. This is subject to some misclassification which is subsequently ignored, whereas the methods I developed took into account the associated uncertainty. In particular, if the semi-Markov model was applied to meal data, we would end up with the same inter-meal information, but the log-normally distributed intra-meal non-feeding periods would be combined with the exponentially distributed feeding events, to give some other distribution of overall meal durations. Hence the model as developed contains all the information that modelling the meal data would have contained, but includes much more information besides. As the main objective is to explore the underlying biological mechanisms that generate the data, it makes sense to retain as much information as possible for model fitting. The main point of hidden Markov models is their ability to have multiple behaviours in states, and so for the modelling of meal data there would be no justification in using these models. For the latent Gaussian model we saw that the model gave similar results when applied to meal data.

Finally, a particular objective for modelling these datasets, with any of the models, was to summarise large datasets by a small number of parameters in order to allow easy comparison of different animals or different groups of animals. Any of the models considered are capable of this and so given a particular dataset, any of the classes of model can be chosen as being the most suitable for that data, and the resulting parameter estimates used to summarise the data.

8.2 Further work

This is divided into two sections — further analysis of the cow feeding data, and issues relating to statistical methodology.

8.2.1 Cow feeding dataset

It has been demonstrated that the use of appropriate models allows large datasets to be summarised by relatively few parameters. This allows differences between individuals to be examined and hopefully related to observed features, and also allows the comparison of groups of animals on different treatments or in different groups or experiments. The full cow feeding dataset consisted of data on three groups of cows — high protein, low protein and choice, and therefore an obvious thing to do would be to compare the three groups in terms of the parameters estimated for the fitted models. This thesis has concentrated on the motivation for the different models and the statistical methodology involved in fitting them, using mainly the eight cows on the high-protein diet for illustration. Actual comparison of the groups has not featured here, being of a descriptive and comparative nature rather than raising new statistical issues. As the thirty day period considered here was part of a longer experiment, it might also be possible to compare the data considered here with similar groups of animals at other times of year, in different stages of lactation or under differing management regimes.

A useful extension of methodology for this dataset would be in the area of multivariate processes, so enabling a whole group or herd of cows to be modelled simultaneously, also with the inclusion of covariates. In this framework it would be possible to impose structure on the model, so that if a given effect was similar for all animals, it could be estimated overall rather than separately for each individual. The latent Gaussian model might be the most obvious model in which this could be done, though we have also seen how the semi-Markov and hidden-Markov models can incorporate covariates.

8.2.2 Statistical methodology

With respect to the latent Gaussian model, there are many areas for further work. The simulation study could be extended to other classes of ARMA model, with the hope that some broad general recommendations could be made with regards to the relative efficiency of methods and the relationships between the

number of lags to be retained and the form of the autocorrelation structure. I have talked about general censored ARMA processes, but have really restricted investigations to simple thresholding only. Other types of censoring remain to be fully investigated. For example, rainfall and solar radiation data was considered by Glasbey et al. (1998). For the rainfall data, zero rainfall corresponds to a censored value below the threshold, whilst above it, the value of the variable is observed. For the solar radiation data, data are missing when the sun is below the horizon. In these cases, when the continuous variable is partially observed, it is required to estimate the variance of the Gaussian variable, whereas in our case the variable was unobservable and hence we could work with standardised variables.

More general categorical data also remains to be investigated, i.e. more than two categories of behaviour, either ordered or unordered. One option would be to have a single latent variable and more than one threshold. Particular ranges of the continuous variable would then correspond to the different behaviours. Clearly this puts restrictions on the order in which behaviours can occur and, unless this was a particular feature of the data being modelled, would be unlikely to allow a sufficiently general description. A less restrictive alternative would be a model with more than one latent variable. McFadden (1982) and Bartholomew and Knott (1999, page 121-122) consider econometric models for which a random variable is associated with each category, and the current category corresponds to the latent variable with the largest current value. This has the potential for modelling behaviour data with many categories.

Hidden semi-Markov models are worth highlighting as an area for future work. A hidden semi-Markov model combines the strengths of both semi-Markov and hidden Markov models. The reason I did not fully explore them here was because of the prohibitive level of the computational effort required. However there may be potential for reducing this in some way and therefore making this type of model more attainable to the general modelling community.

Finally, model choice for non-nested models has been discussed at various stages throughout this project. Use was made of AIC and BIC, recognising that there were some flaws in their application, but nevertheless use of them in conjunction with other ideas showed them to be useful. A parametric bootstrap approach was considered, showing itself to be a useful technique, producing the expected results. However it would be desirable to have more formal results and more rigid methodology, especially for the comparison of more than two models simultaneously. This is not an issue which appears to have been addressed before and

suggestions were made in the last chapter in terms of using Mahalanobis distance or results for order statistics in more than one dimension. This would be a useful extension to existing methodology.

Appendix A

Data

This appendix displays some of the cow feeding data introduced in Section 1.2 and used throughout the thesis. The full dataset available consisted of thirty days of data for a total of 34 cows; eight were on a high-protein (HP) diet, ten were on a low-protein diet (LP) and the remaining sixteen had access to both feed types (CH). Here I display data for a subset of the animals in order to illustrate the similarities and differences between individuals. Most of the models in the thesis are illustrated with data from the high-protein cows and therefore data are displayed for all eight high-protein cows in Figures A.1–A.8. A few animals from the other two feeding regimes are also shown; data from four of the low-protein (LP) animals, Cows 9, 75, 118 and 224, are shown in Figures A.9–A.12, and four of the cows on the choice (CH) diet, Cows 43, 76, 132 and 165, are shown in Figures A.13–A.16. In all cases the data are shown as binary time series, raised values of the signal denoting feeding periods. For the cows on the choice diet, the height of the signal denotes which type of food was being eaten during that visit, i.e. full height for the high-protein and half height for the low-protein. Days are numbered from 106 to 135, corresponding to 16 April – 15 May 1995, i.e. days are numbered from 1 January 1995 being Day 1. For the purposes of display, days start at 08:00, the time at which the electronic timers on the feeders were daily reset.

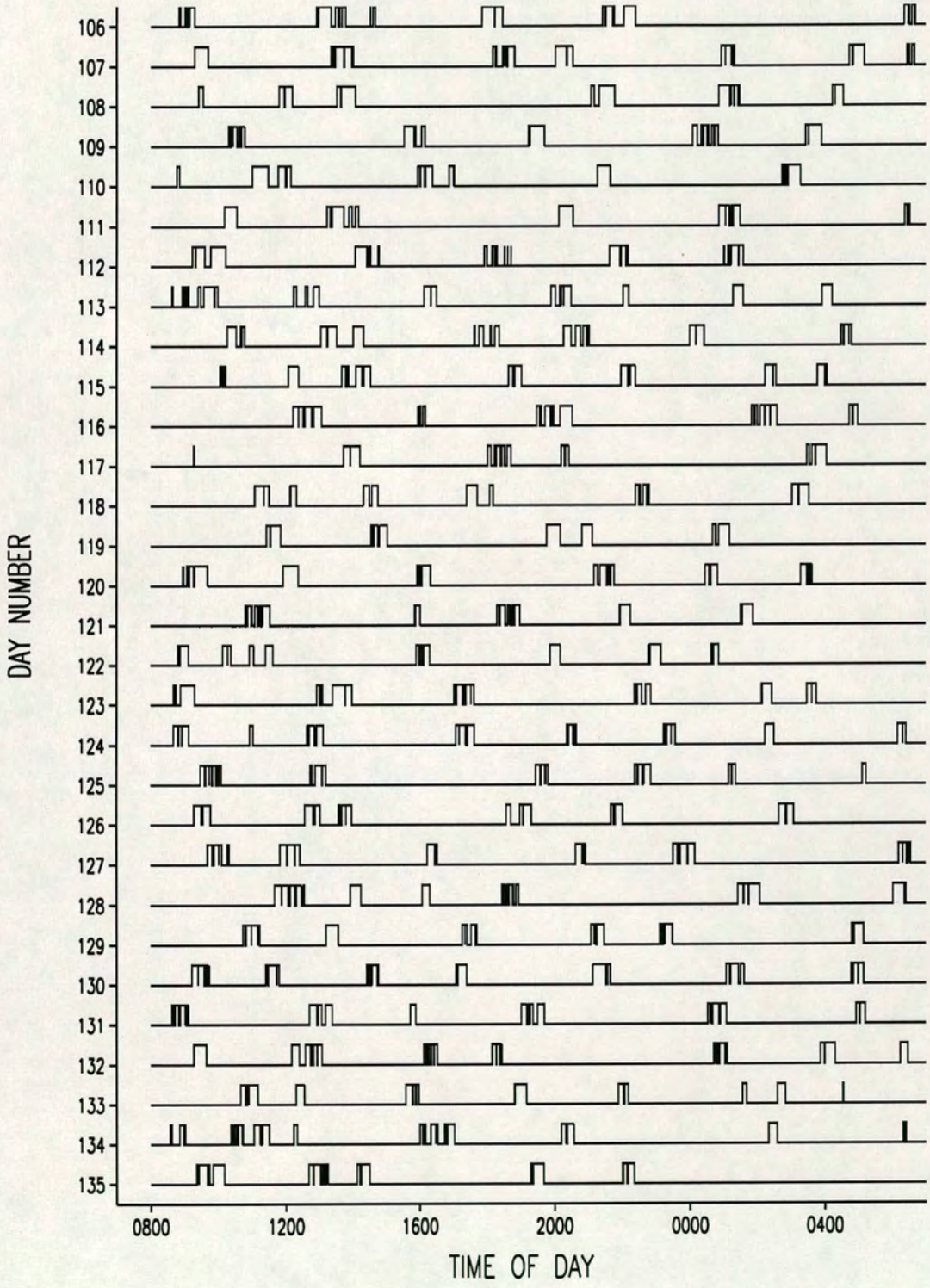

Figure A.1: *Feeding data for Cow 5 (HP).*

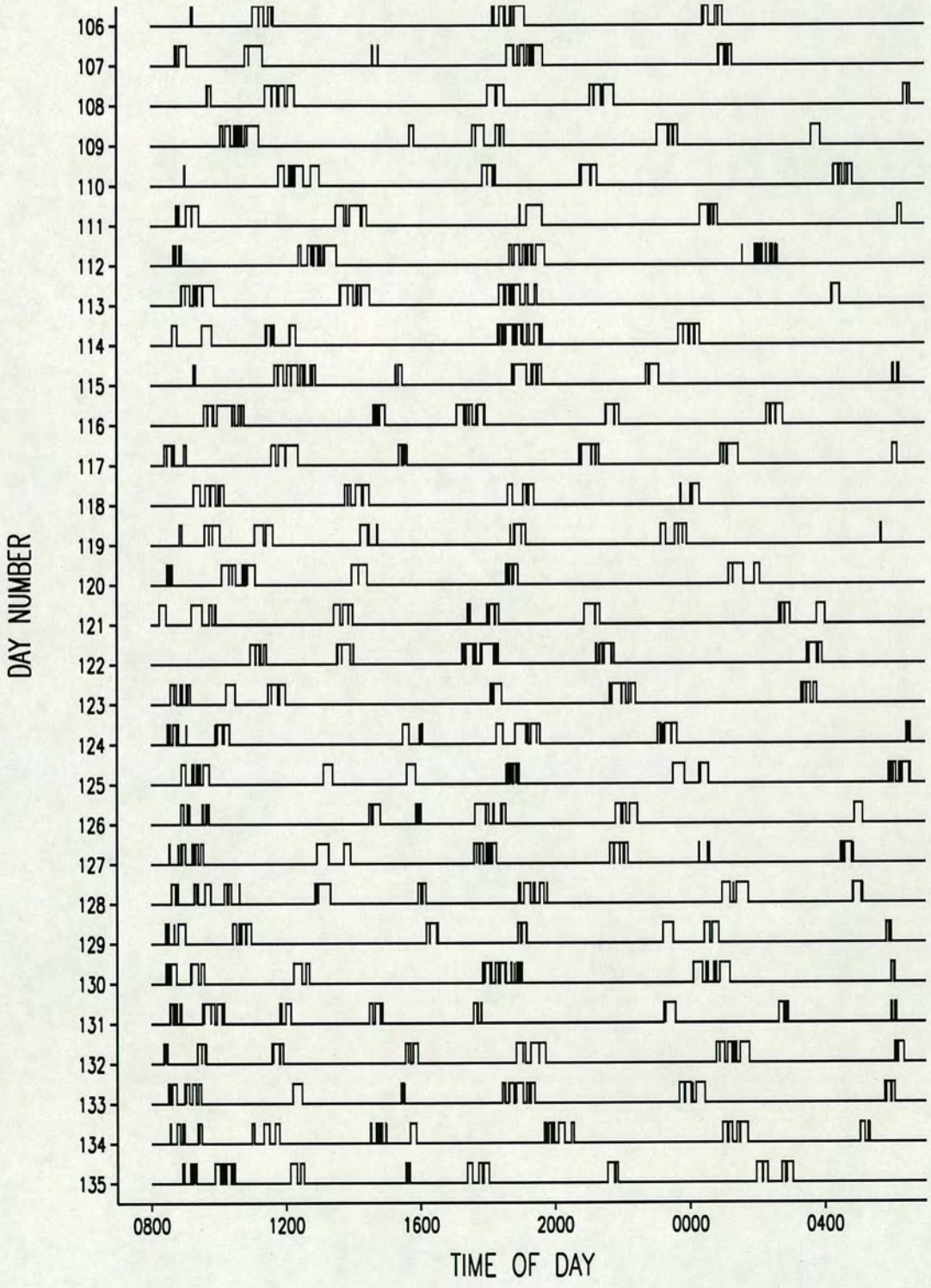

Figure A.2: *Feeding data for Cow 41 (HP).*

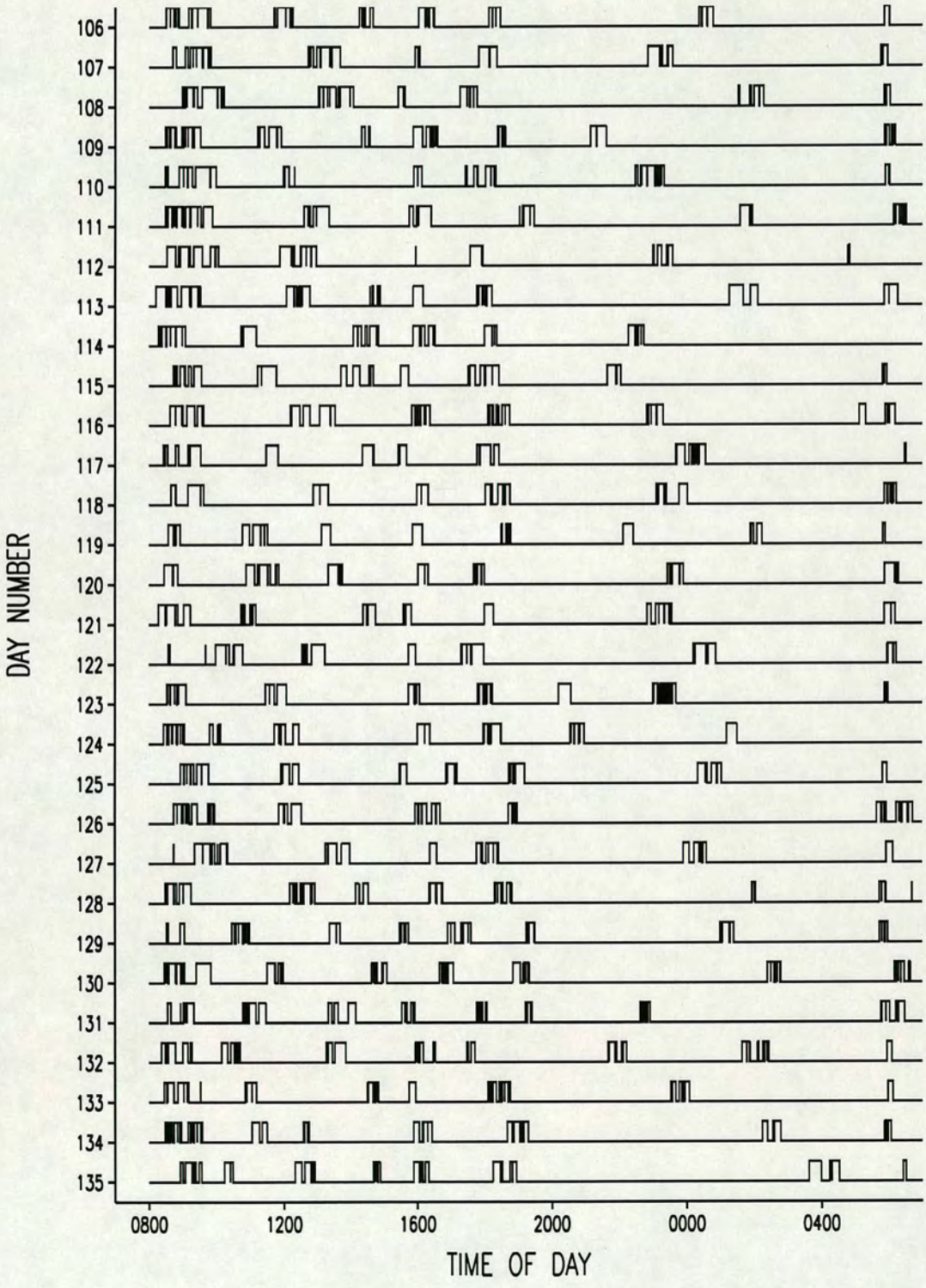

Figure A.3: *Feeding data for Cow 108 (HP).*

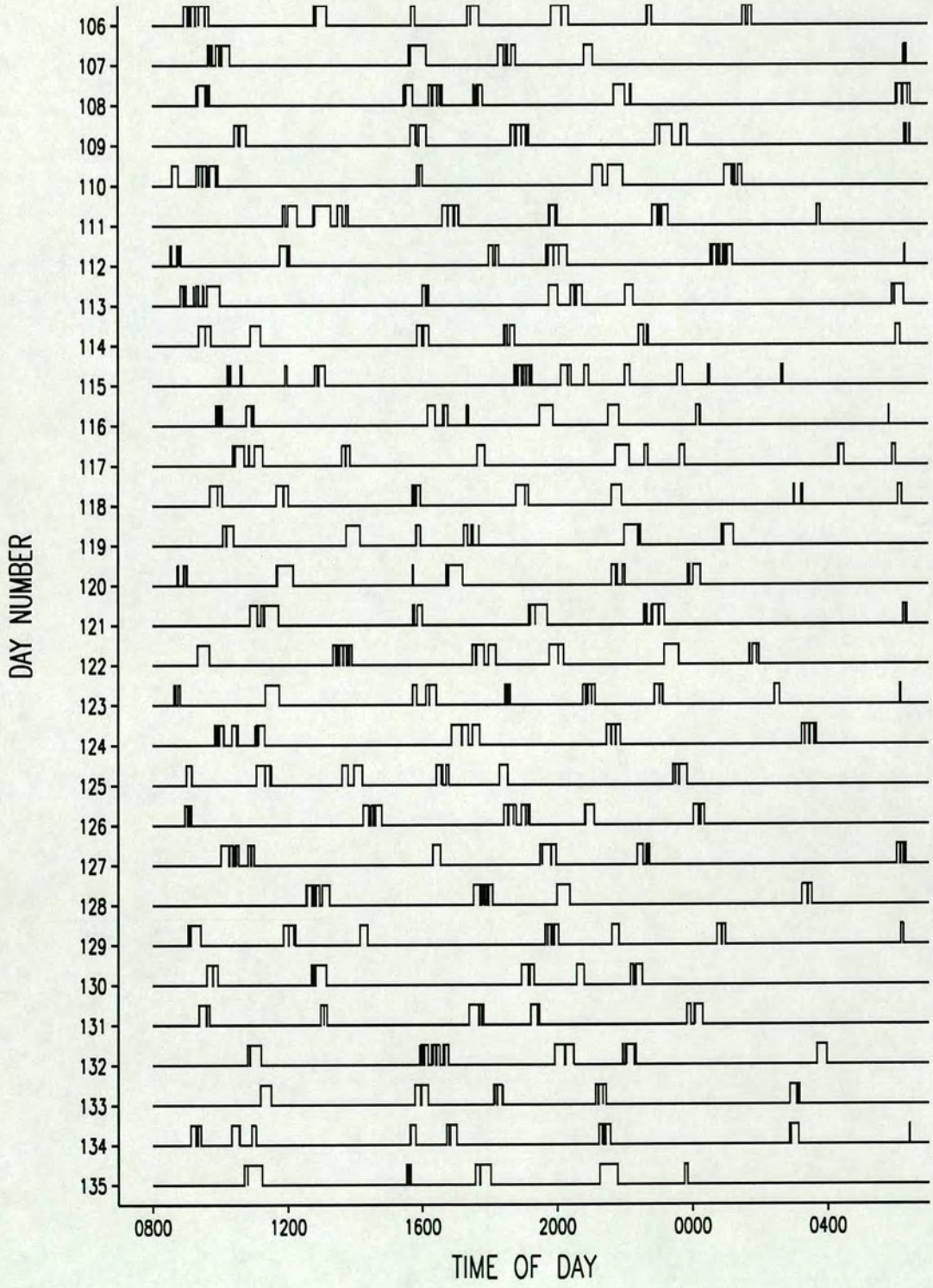

Figure A.4: *Feeding data for Cow 169 (HP).*

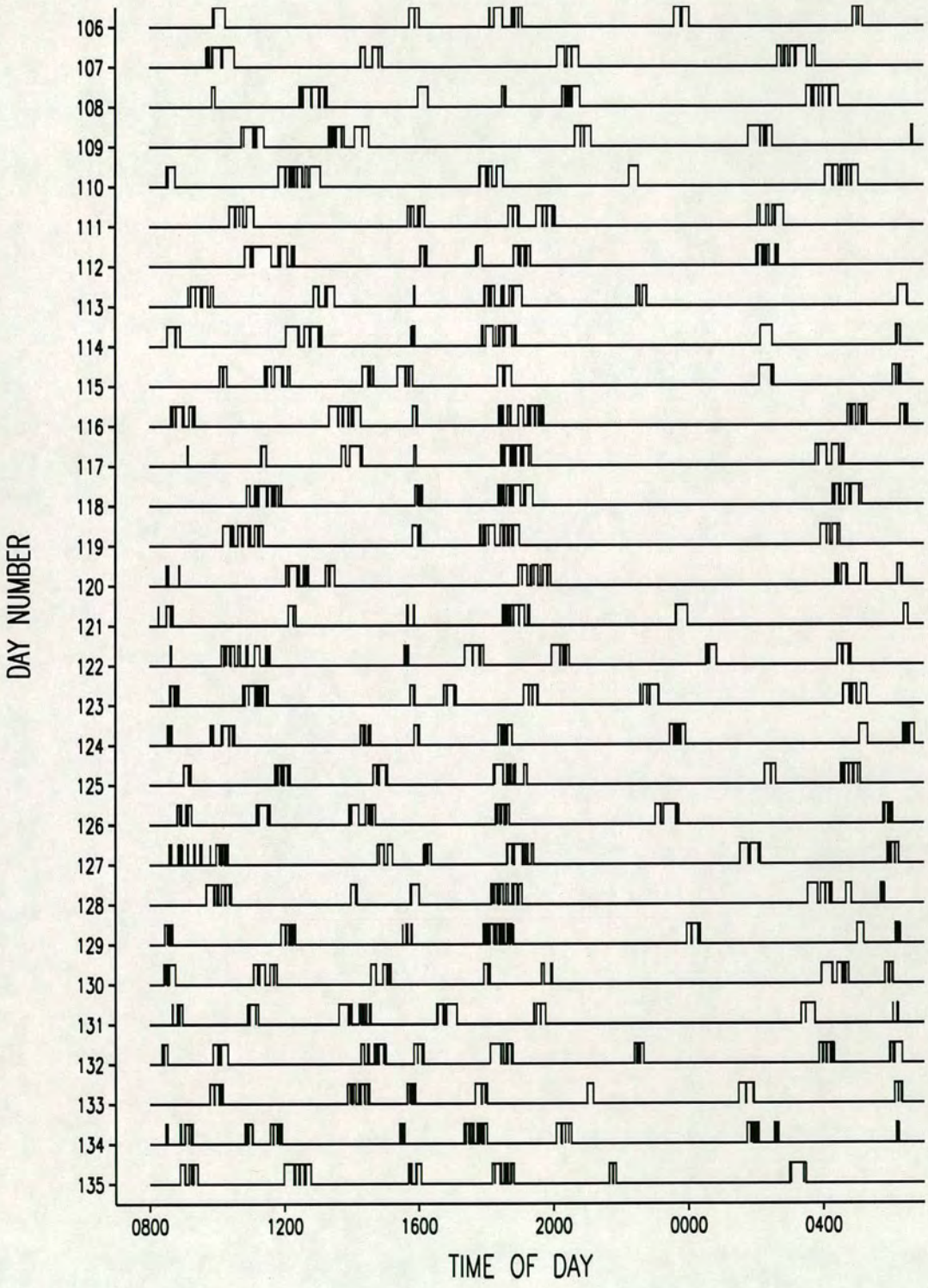

Figure A.5: *Feeding data for Cow 170 (HP).*

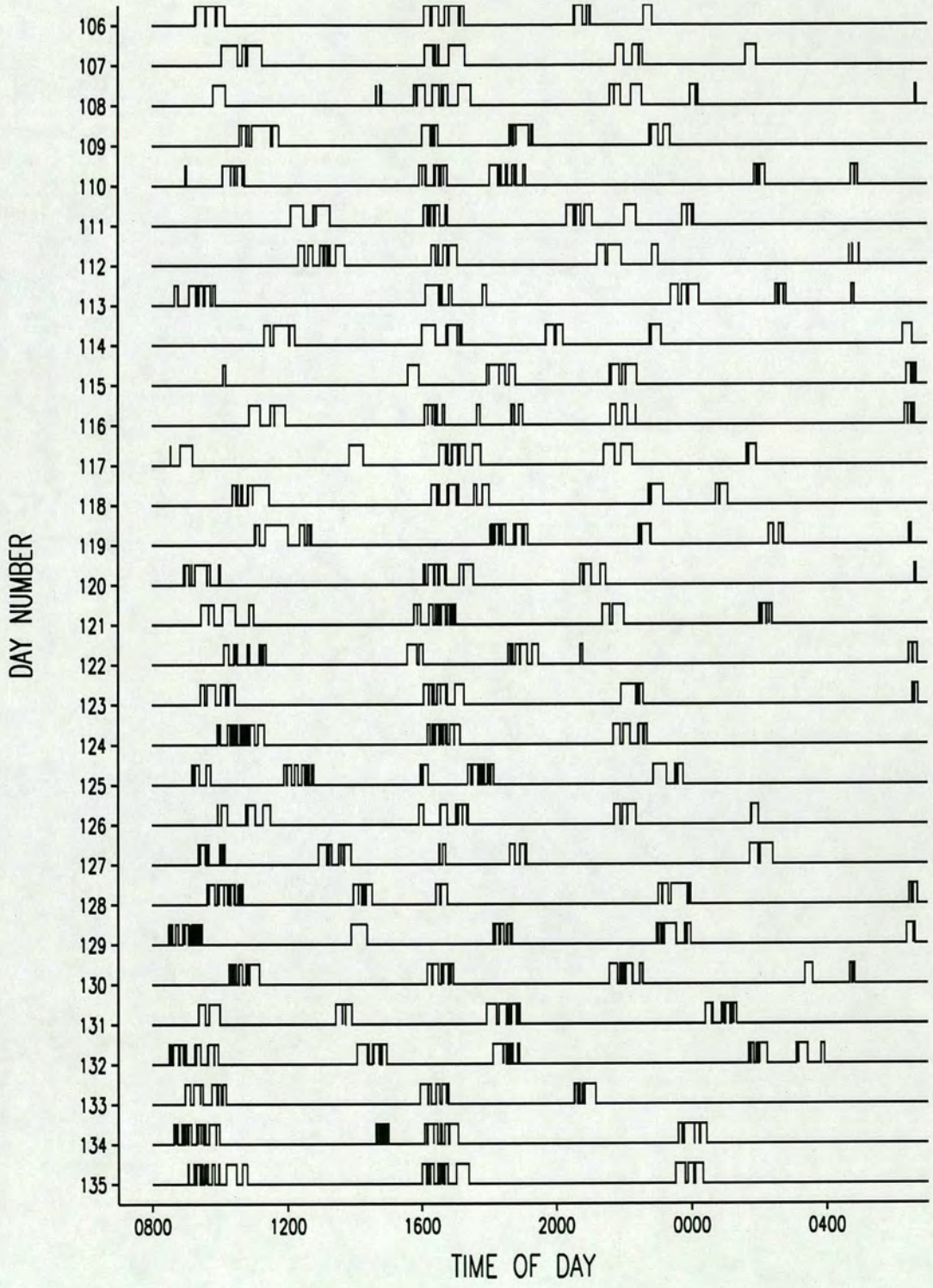

Figure A.6: *Feeding data for Cow 182 (HP).*

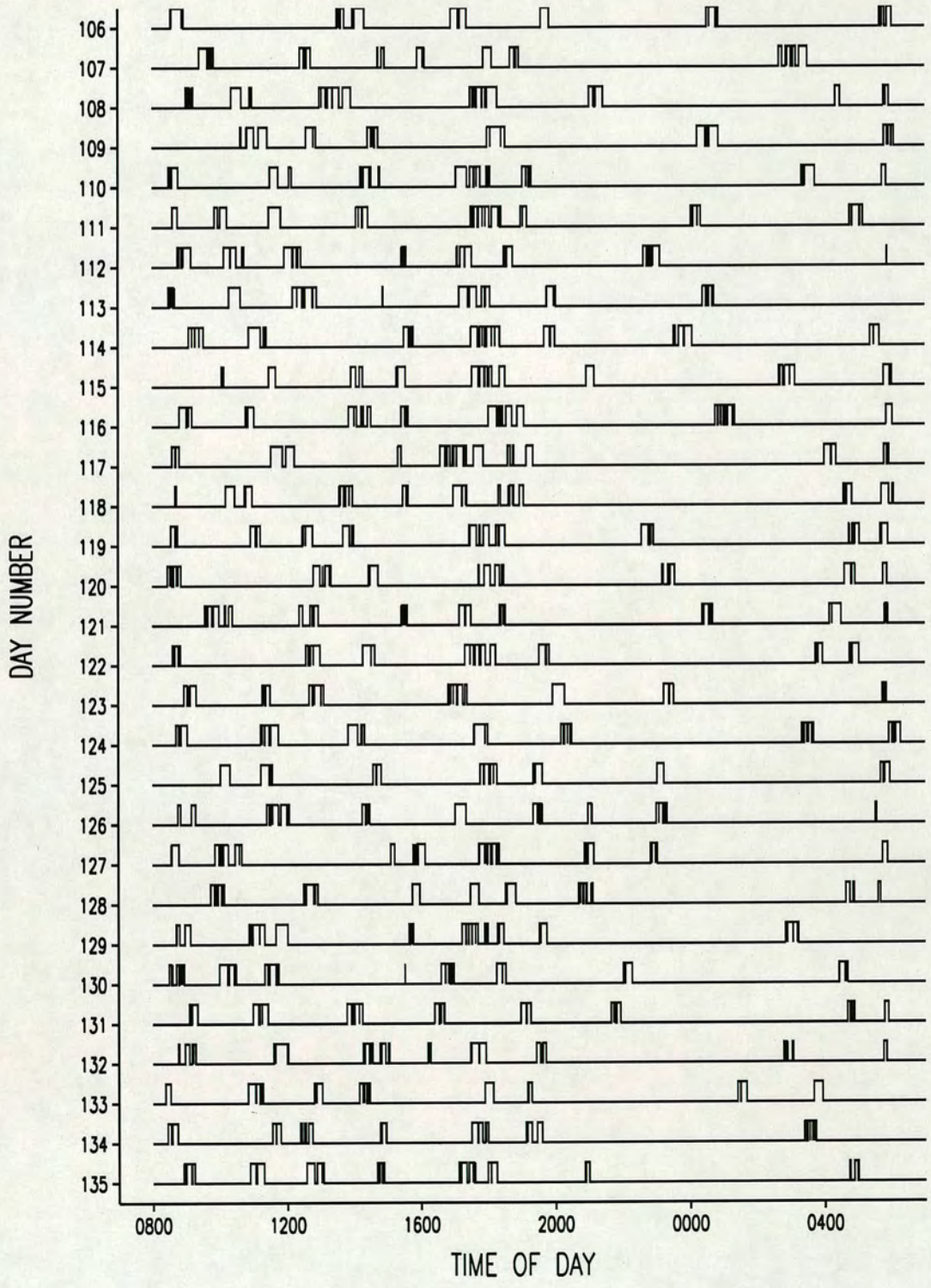

Figure A.7: *Feeding data for Cow 194 (HP).*

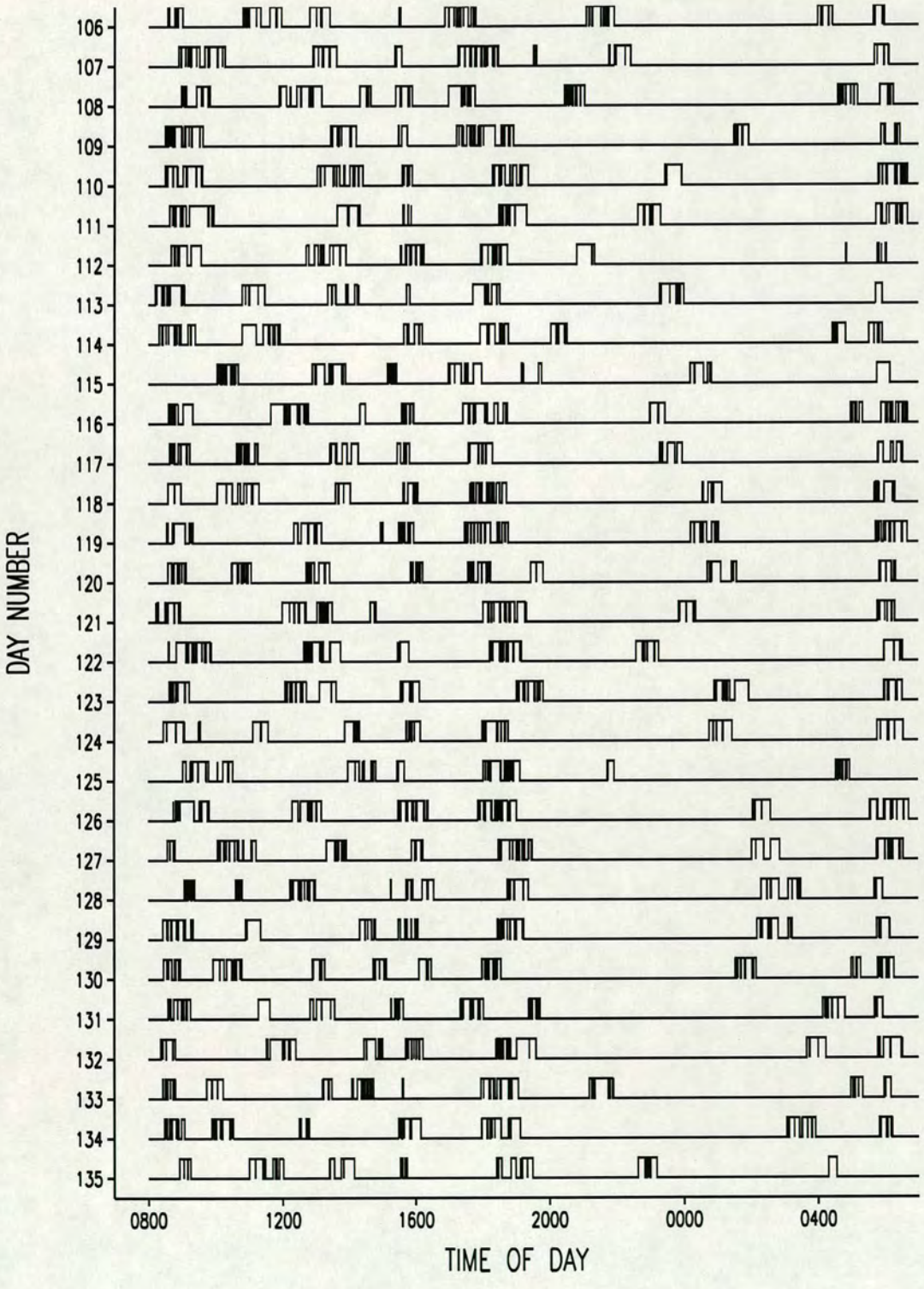

Figure A.8: *Feeding data for Cow 221 (HP).*

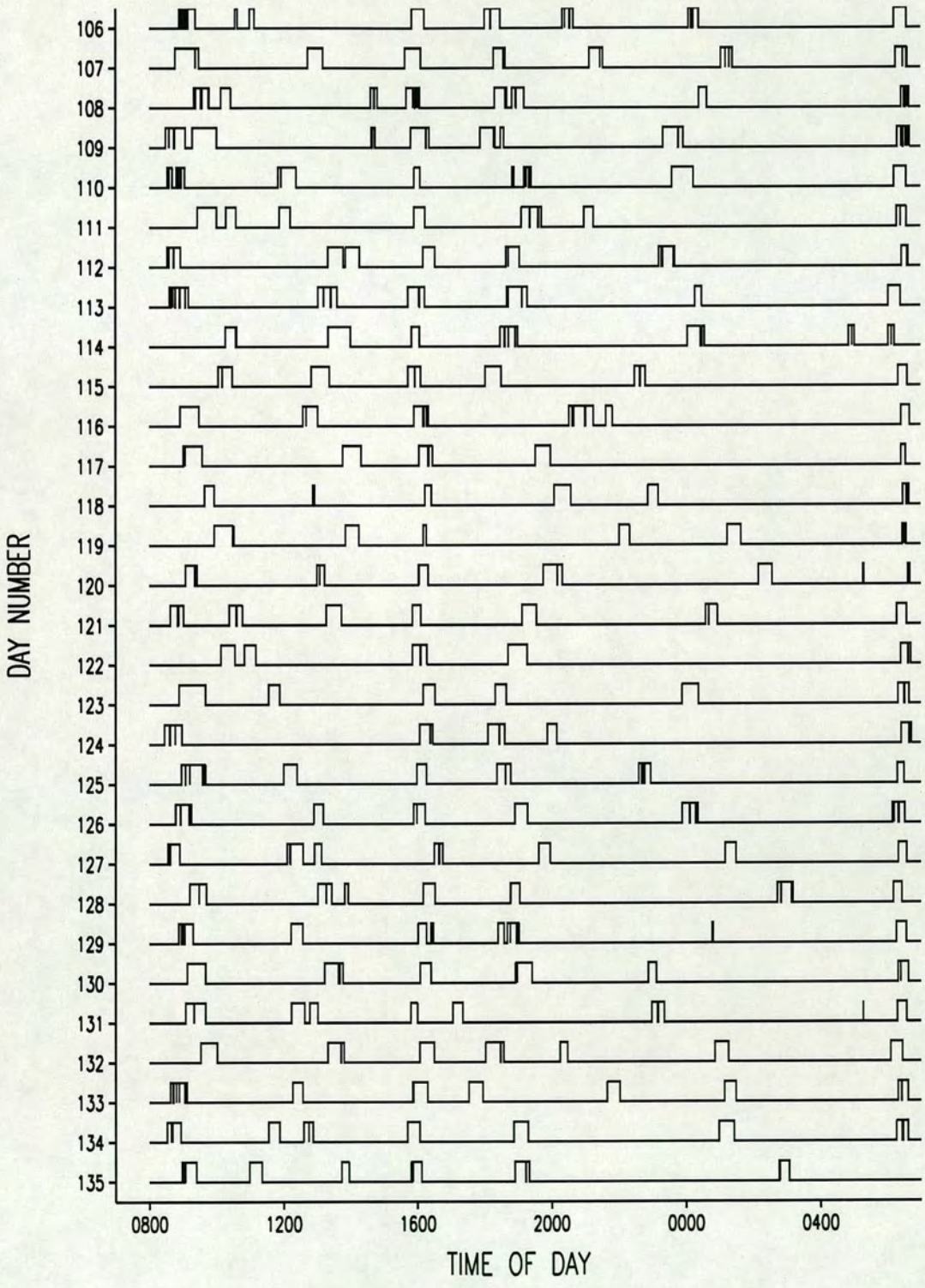

Figure A.9: *Feeding data for Cow 9. (LP)*

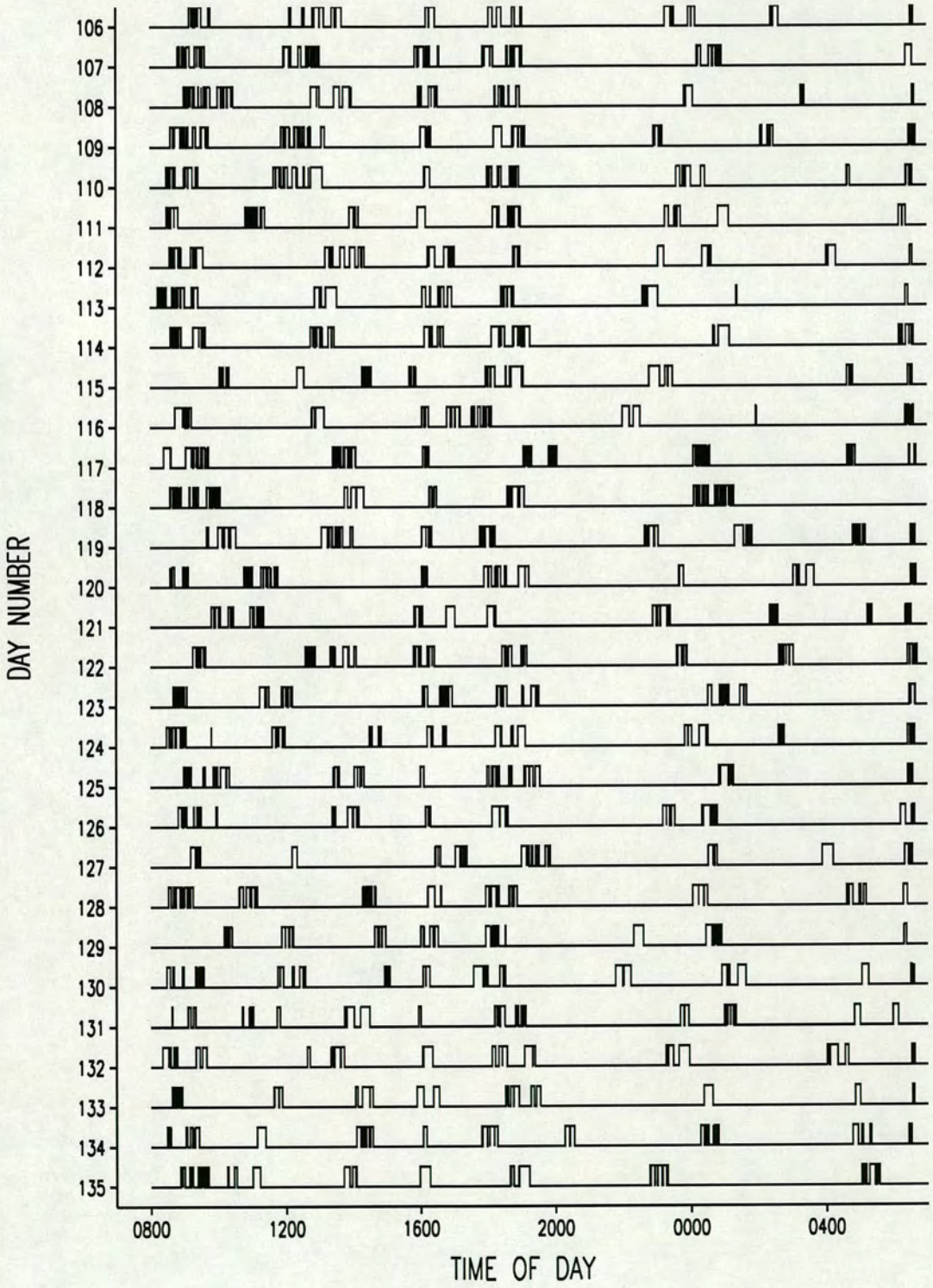

Figure A.10: *Feeding data for Cow 75. (LP)*

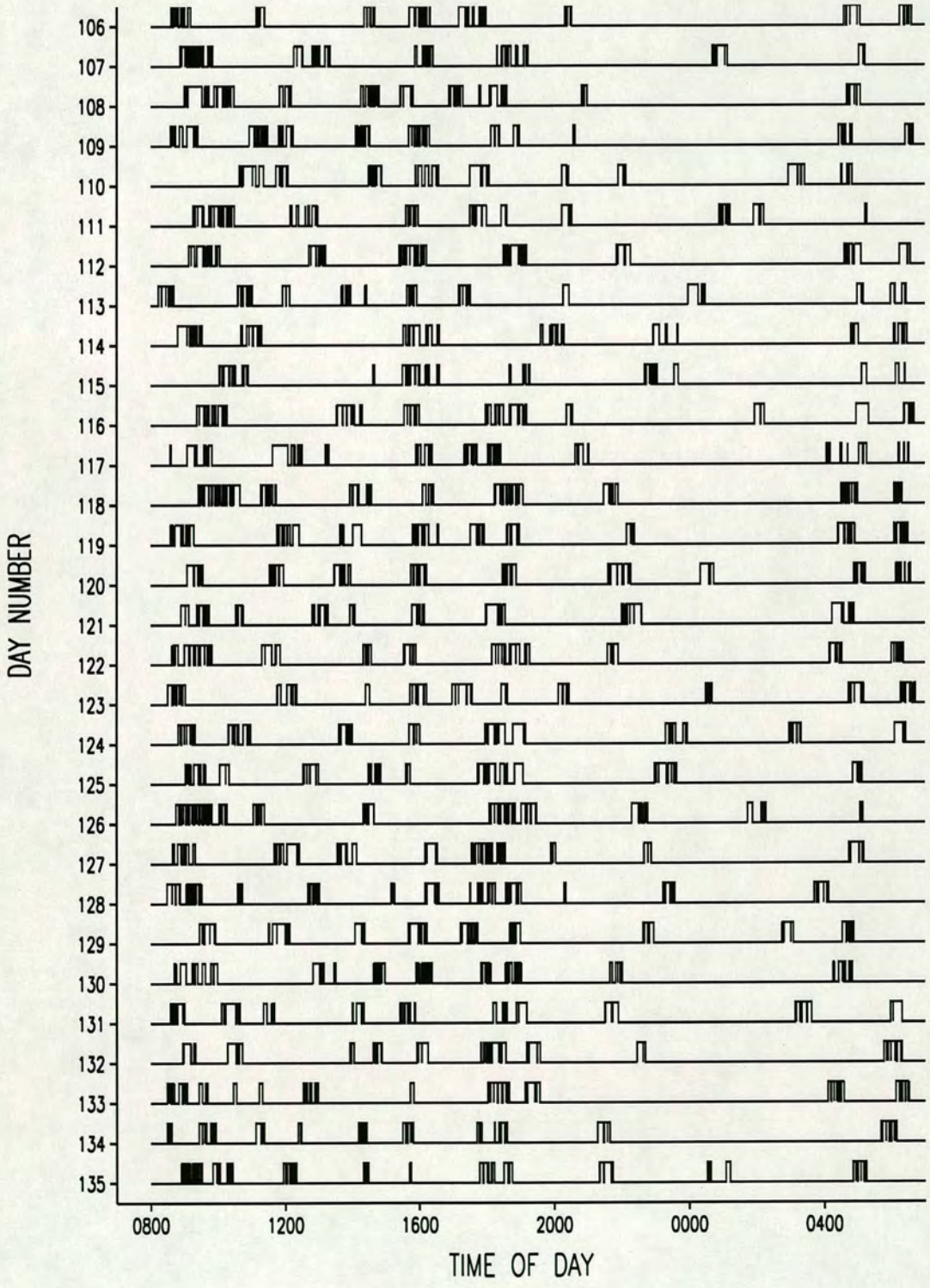

Figure A.11: *Feeding data for Cow 118. (LP)*

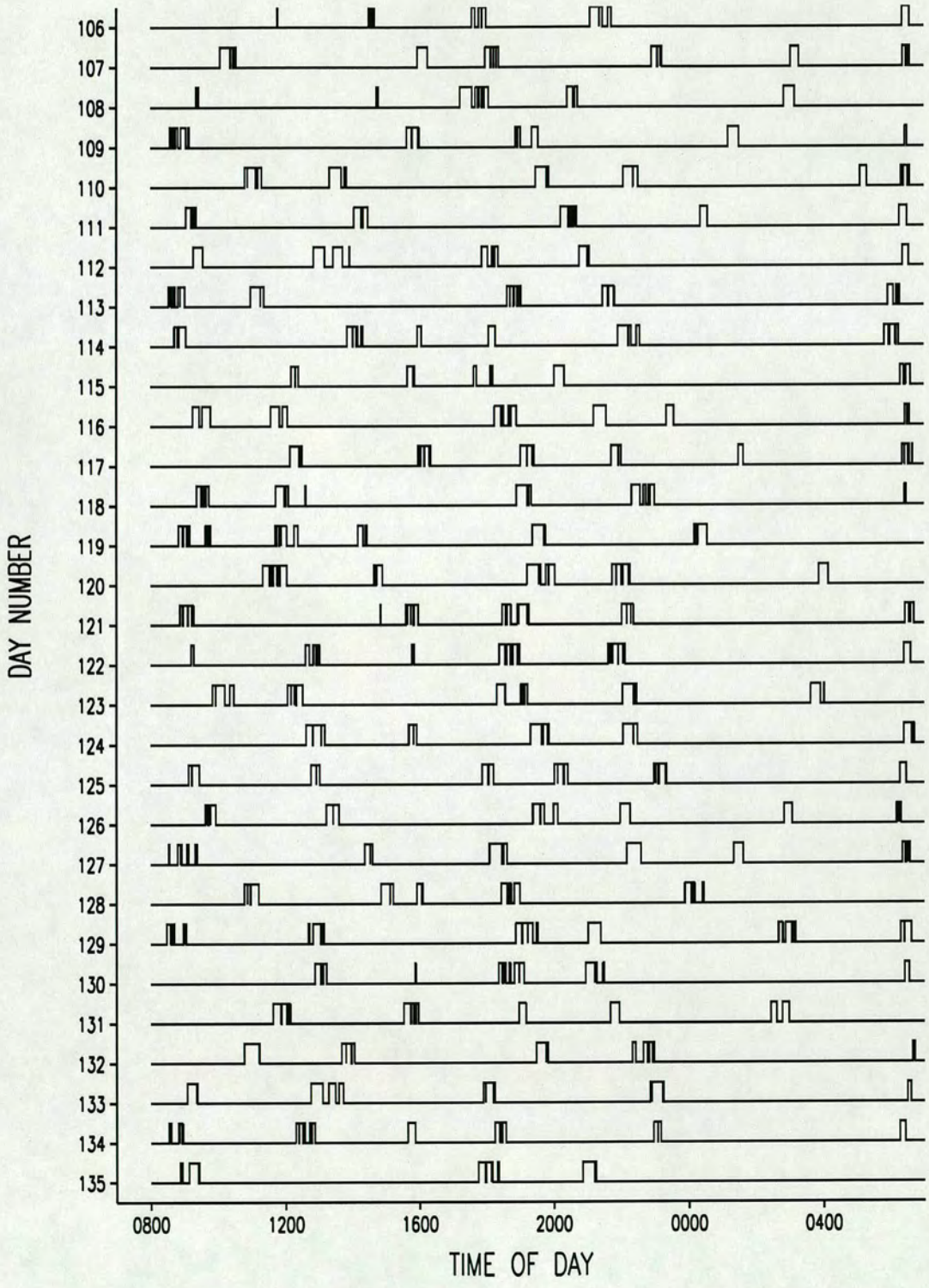

Figure A.12: *Feeding data for Cow 224. (LP)*

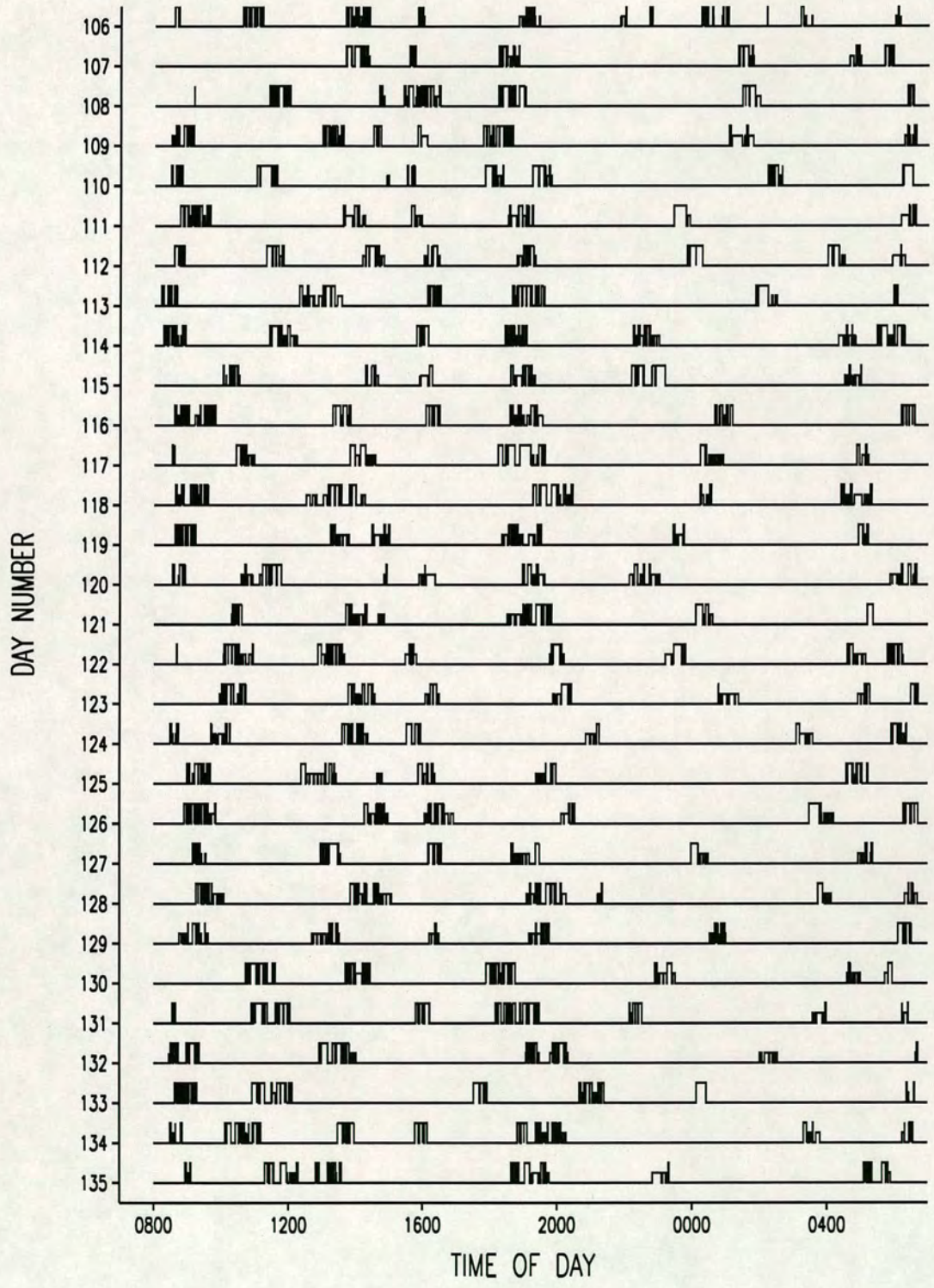

Figure A.13: *Feeding data for Cow 43. (CH)*

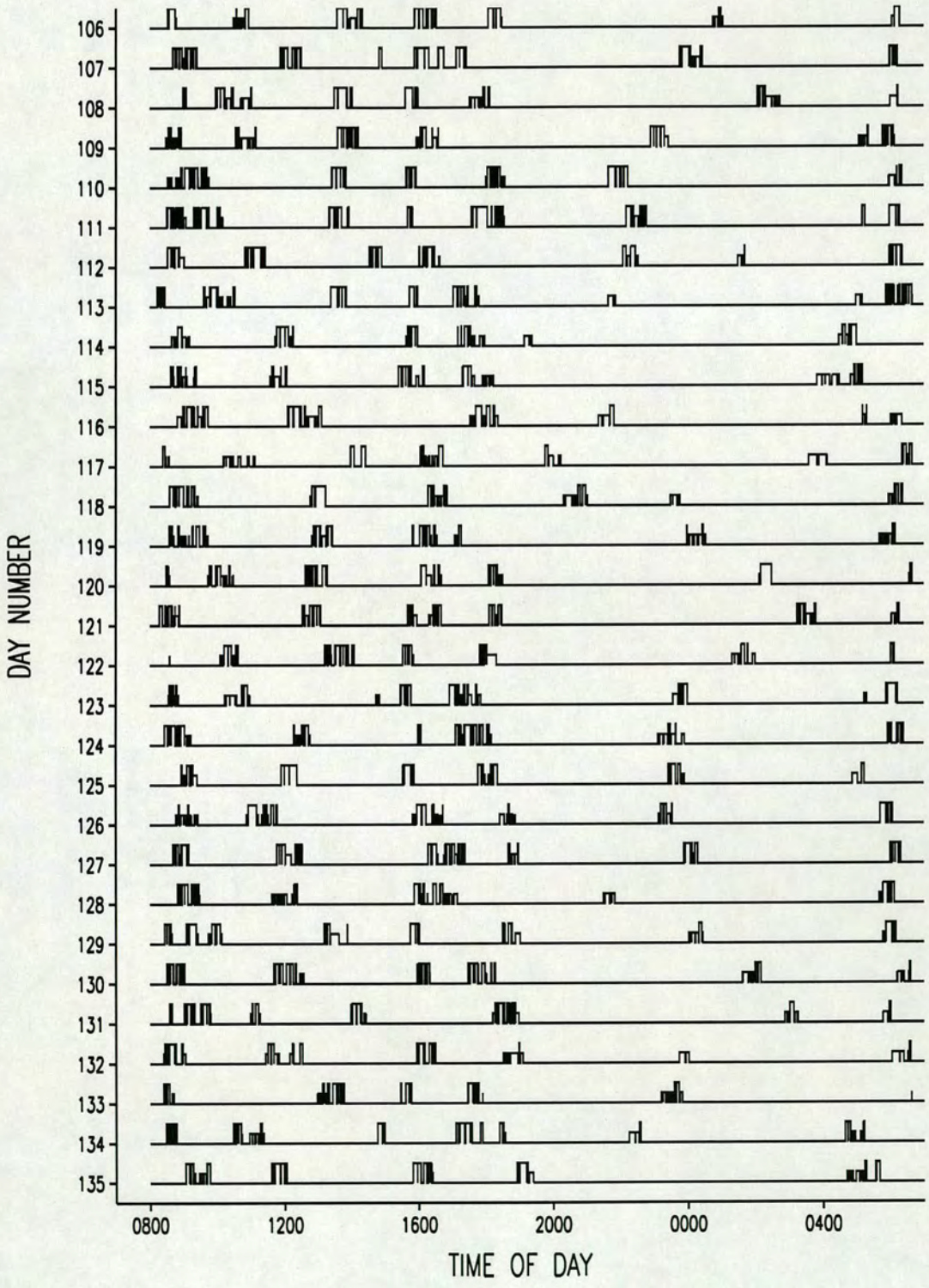

Figure A.14: *Feeding data for Cow 76. (CH)*

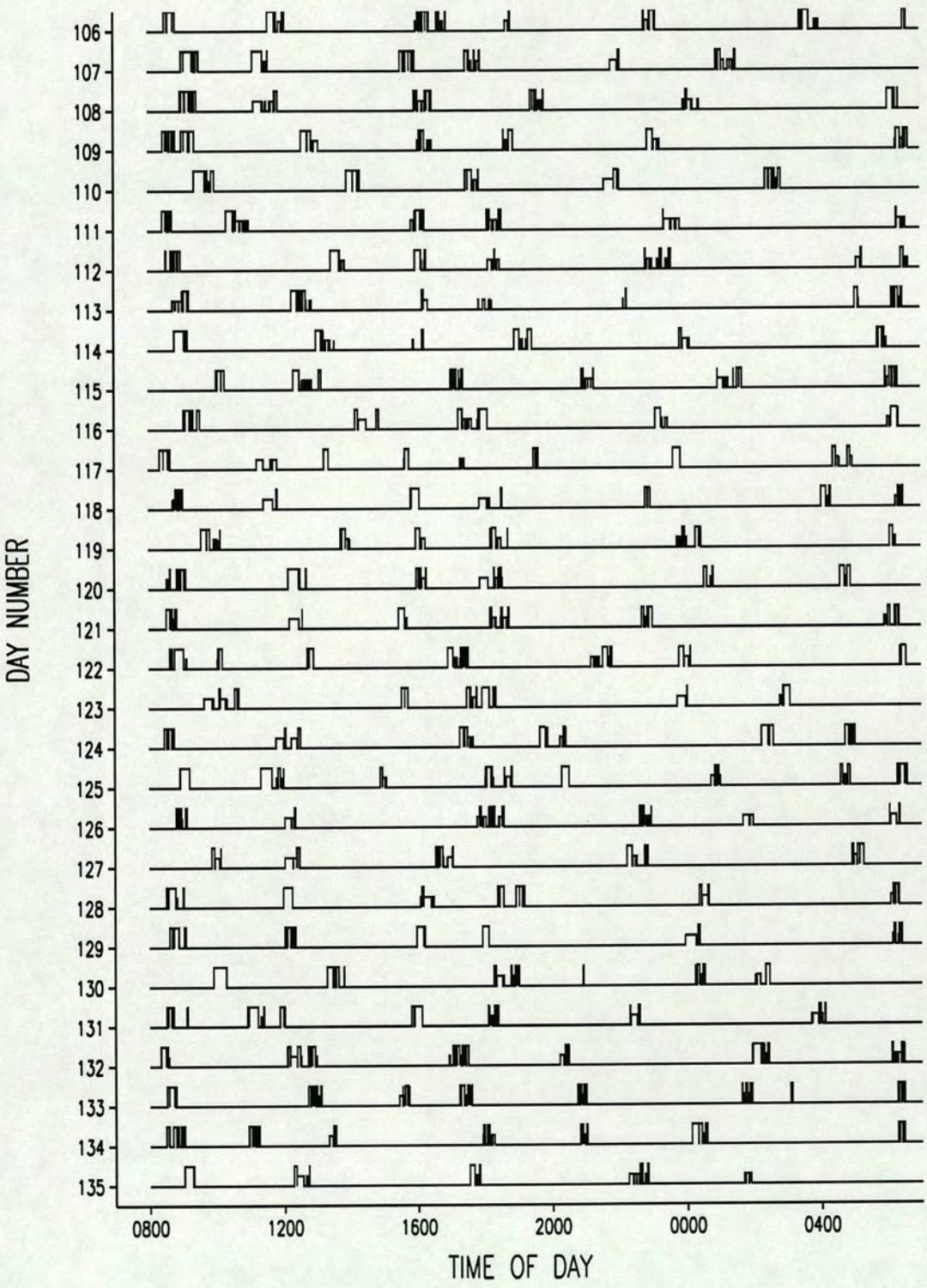

Figure A.15: *Feeding data for Cow 132. (CH)*

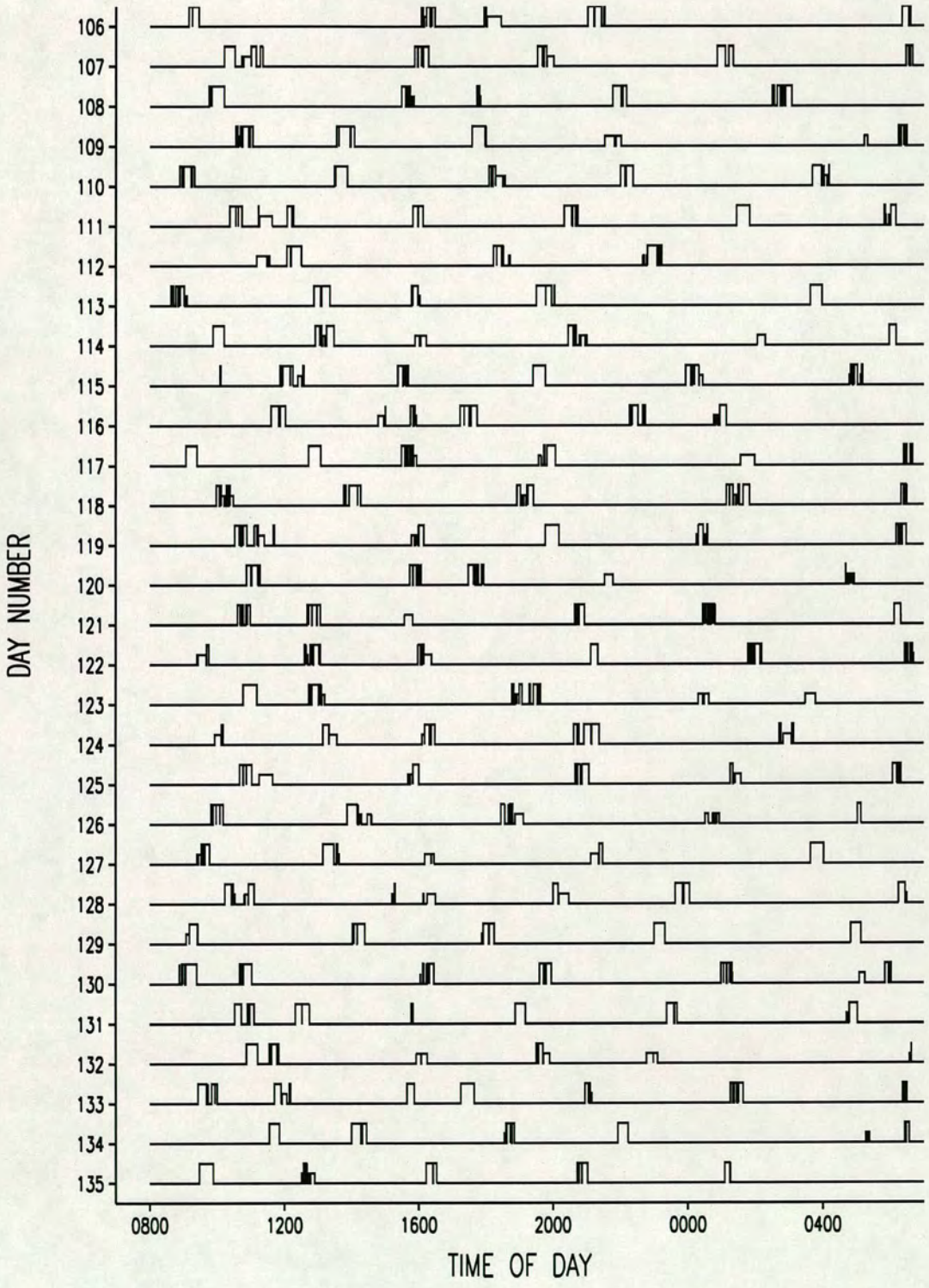

Figure A.16: *Feeding data for Cow 165. (CH)*

Appendix B

Example CODA output

CODA, properly called *Convergence Diagnostics and Output Analysis Software for Gibbs sampling output* (Best et al., 1995, 1997), is a package designed to assess the convergence of Markov chains produced by MCMC methods, specifically those using Gibbs sampling. It calculates summary statistics and produces plots of the chains; it also performs more formal tests to check for convergence.

Here we present some example CODA output as used to check the convergence of the Markov chains produced using the MCMC methods of Section 3.5 in the simulation study described in Section 3.6. CODA output was produced for all series in the simulation study; here the idea is not to present or even summarise all the results obtained, merely to describe and illustrate the plots and methods that were used.

In Section B.1, I briefly describe the output that was used to assess convergence, namely

- summary statistics,
- plots of each chain, with kernel density plots,
- results of the Geweke diagnostic,
- results of the Heidelberger and Welch test.

Section B.2 then gives example output for four realisations of each of

- AR(1) process with $\phi = 0.6$, thresholding at 1 standard deviation and series length 1000,
- MA(1) process with $\theta = -0.3$, thresholding at 0 and series length 100.

B.1 Description of output

B.1.1 Summary statistics and plots

Summary statistics produced by CODA include the empirical mean and the standard deviation, which estimates the square root of the variance of the posterior distribution. Several standard errors of the mean are also given — the naive estimate assumes the estimates to be independent; the time-series estimate is calculated from the spectral density estimate; the batch estimate is calculated after the sample has been divided into batches of size 25, with the hope that batches are reasonably independent, implying the lag 1 autocorrelation given should be close to zero. Quantiles are also presented.

Plots show the trace of the Markov chain, in our case showing estimates of the parameters from all 10500 iterations. For the kernel density plots of the posterior distributions, only the last 10000 iterations are used.

B.1.2 Convergence diagnostics

Both the convergence diagnostics considered here have been developed for single chains and are suitable when the statistic of interest is the mean. They are fully reviewed in Brooks and Roberts (1998), enough description being given here only to be sufficient to interpret the output presented later.

Geweke's diagnostic compares the first 10% of the chain with the last 50%, involving calculation of the sample mean and asymptotic variance in each window and performing a z-test. Hence the values printed are compared with the standard normal distribution to check for evidence against convergence. The plots are produced by splitting the chain into 50 segments, Geweke's diagnostic being computed and plotted for each segment. A large number of z-scores falling outside the horizontal lines plotted at ± 1.96 would suggest possible convergence failure.

The diagnostic due to Heidelberger and Welch is based on 'Brownian bridge' theory and uses the Cramer-von-Mises statistic to test the null hypothesis that the chain forms a stationary process. It also performs a halfwidth test by using the sample mean and its asymptotic standard error to calculate the halfwidth of the confidence interval for the mean. The default test is passed if the halfwidth is less than 0.1 times the sample mean, indicating that the posterior mean is estimated with acceptable precision.

B.2 Sample Output

B.2.1 AR(1) process

The following is example output for an AR(1) process, with $\phi = 0.6$, thresholded at 1 standard deviation, for realisations of length 1000.

```
SUMMARY STATISTICS
==================

Iterations used = 501:10500
Thinning interval = 1
Sample size per chain = 10000

Batch size for calculating Batch SE = 25


1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:


Chain: AR6_1t
=============

-+----------+-------------------------------------------------------+-
 | VARIABLE |   Mean      SD        Naive SE  Time-series SE    |
 | ======== |   ====      ==        ========  ==============    |
 |          |                                                     |
 | phi[1]   |   0.576000  0.039400  0.000394  0.001100          |
 | phi[2]   |   0.591000  0.040700  0.000407  0.001220          |
 | phi[3]   |   0.585000  0.042600  0.000426  0.001290          |
 | phi[4]   |   0.614000  0.039100  0.000391  0.001120          |
 |          |                                                     |
-+----------+-------------------------------------------------------+-
-+----------+-------------------------------------+-
 | VARIABLE |   Batch SE  Lag-1 batch autocorr    |
 | ======== |   ========  ====================    |
 |          |                                     |
 | phi[1]   |   0.001050  0.173000                |
 | phi[2]   |   0.001150  0.145000                |
 | phi[3]   |   0.001260  0.230000                |
 | phi[4]   |   0.001150  0.170000                |
 |          |                                     |
-+----------+-------------------------------------+-
```

2. Quantiles for each variable:

Chain: AR6_1t
=============

VARIABLE	2.5%	50%	97.5%
phi[1]	0.497	0.577	0.652
phi[2]	0.508	0.591	0.668
phi[3]	0.500	0.586	0.667
phi[4]	0.538	0.614	0.689

GEWEKE CONVERGENCE DIAGNOSTIC (Z-score):
==

Iterations used = 501:10500
Thinning interval = 1
Sample size per chain = 10000

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

VARIABLE	AR6_1t
phi[1]	0.682
phi[2]	0.934
phi[3]	1.100
phi[4]	0.351

HEIDELBERGER AND WELCH STATIONARITY AND INTERVAL HALFWIDTH TESTS:
==

Iterations used = 501:10500
Thinning interval = 1
Sample size per chain = 10000

Precision of halfwidth test = 0.1

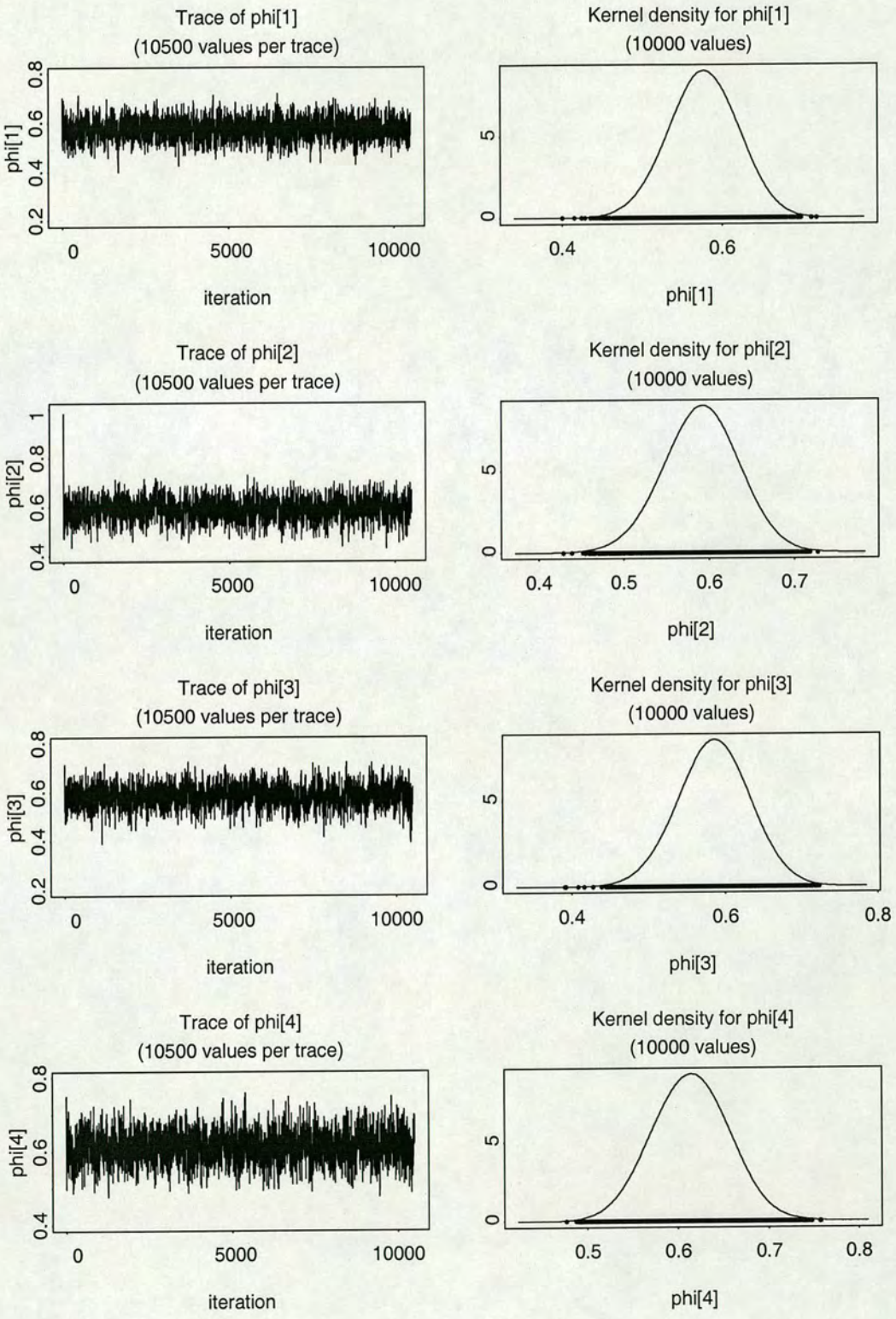

Figure B.1: *Traces of parameter estimates over the 10500 iterations and kernel density plots based on the last 10000 estimates, for each of four runs.*

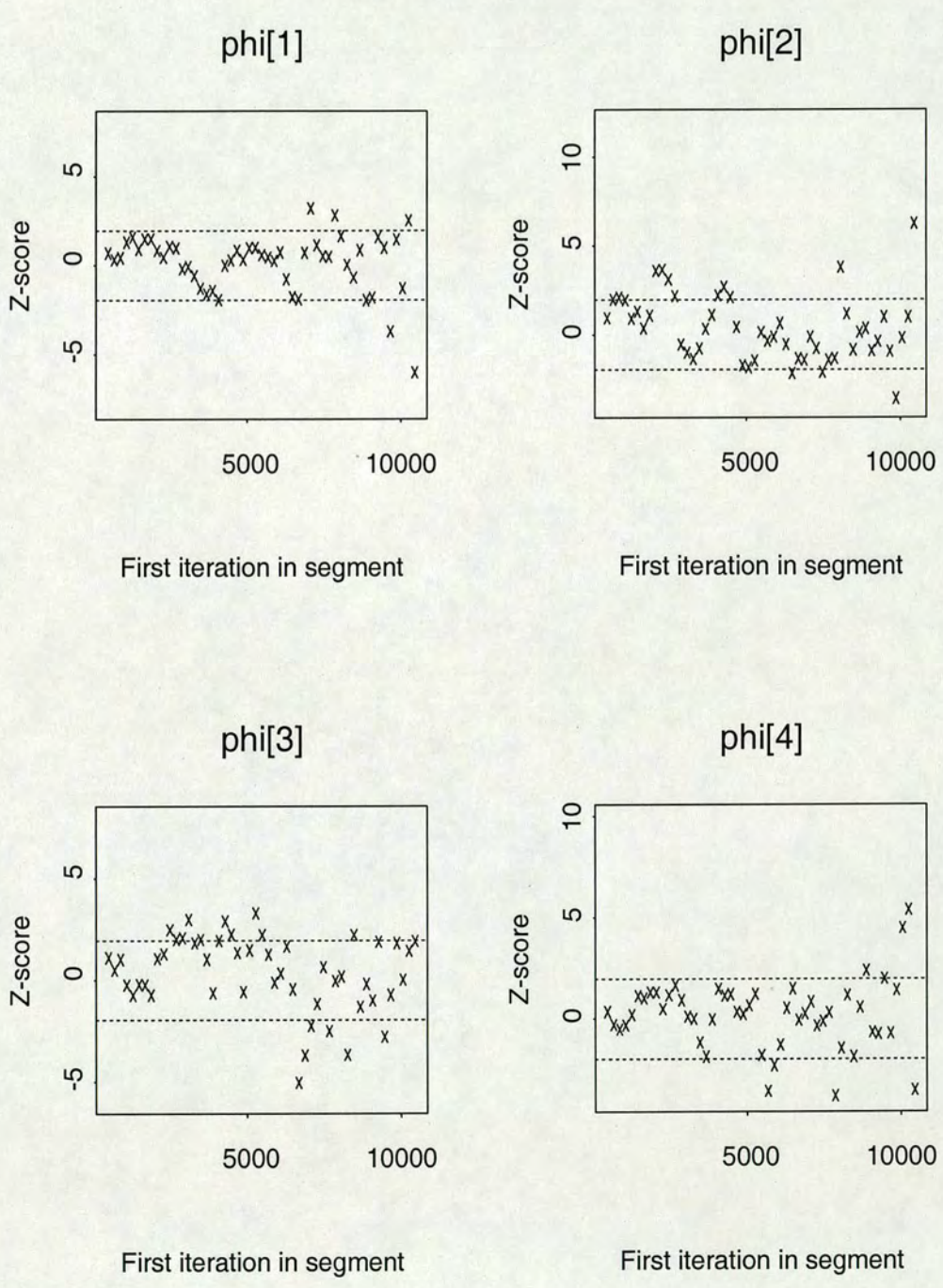

Figure B.2: *Geweke z-scores, for each of four runs.*

```
Chain: AR6_1t
=============

-+----------+-------------------------------------------------------+-
 |          |  Stationarity   # of iters.  # of iters.  C-vonM     |
 | VARIABLE |      test         to keep    to discard    stat.     |
 | ======== |  ============   ==========  ==========  ======     |
 |          |                                                      |
 | phi[1]   |  passed         10000        0           0.0759    |
 | phi[2]   |  passed         10000        0           0.3790    |
 | phi[3]   |  passed         10000        0           0.2380    |
 | phi[4]   |  passed         10000        0           0.0666    |
 |          |                                                      |
-+----------+-------------------------------------------------------+-

-+----------+---------------------------------+-
 |          |  Halfwidth                      |
 | VARIABLE |    test      Mean    Halfwidth  |
 | ======== |  =========   ====    =========  |
 |          |                                 |
 | phi[1]   |  passed      0.576   0.00215    |
 | phi[2]   |  passed      0.591   0.00239    |
 | phi[3]   |  passed      0.585   0.00253    |
 | phi[4]   |  passed      0.614   0.00220    |
 |          |                                 |
-+----------+---------------------------------+-
```

For the four realisations presented, the plots in Figure B.1 look sufficiently stable with relatively little burn-in required; posterior densities look normal. Geweke's diagnostic provides no evidence against convergence, although the plots of Figure B.2 do have a number of points outside the limits ± 1.96. The Heidelberger-Welch tests are all passed. Therefore, apart from the Geweke plots, we have little evidence against convergence.

B.2.2 MA(1) process

The following is example output for an MA(1) process, with $\theta = -0.3$, thresholded at 0 standard deviations, for realisations of length 100.

```
SUMMARY STATISTICS
==================

Iterations used = 501:10500
Thinning interval = 1
Sample size per chain = 10000
```

Batch size for calculating Batch SE = 25

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

Chain: MA3_0h
=============

VARIABLE	Mean	SD	Naive SE	Time-series SE
theta[1]	-0.33500	0.20600	0.00206	0.00798
theta[2]	-0.03060	0.20300	0.00203	0.00671
theta[3]	-0.40600	0.18900	0.00189	0.00775
theta[4]	-0.53900	0.20100	0.00201	0.00855

VARIABLE	Batch SE	Lag-1 batch autocorr
theta[1]	0.00787	0.37000
theta[2]	0.00677	0.11500
theta[3]	0.00750	0.38500
theta[4]	0.00832	0.53500

2. Quantiles for each variable:

Chain: MA3_0h
=============

VARIABLE	2.5%	50%	97.5%
theta[1]	-0.7840	-0.3230	0.0410
theta[2]	-0.4490	-0.0246	0.3540
theta[3]	-0.8220	-0.3930	-0.0681
theta[4]	-0.9540	-0.5270	-0.2040

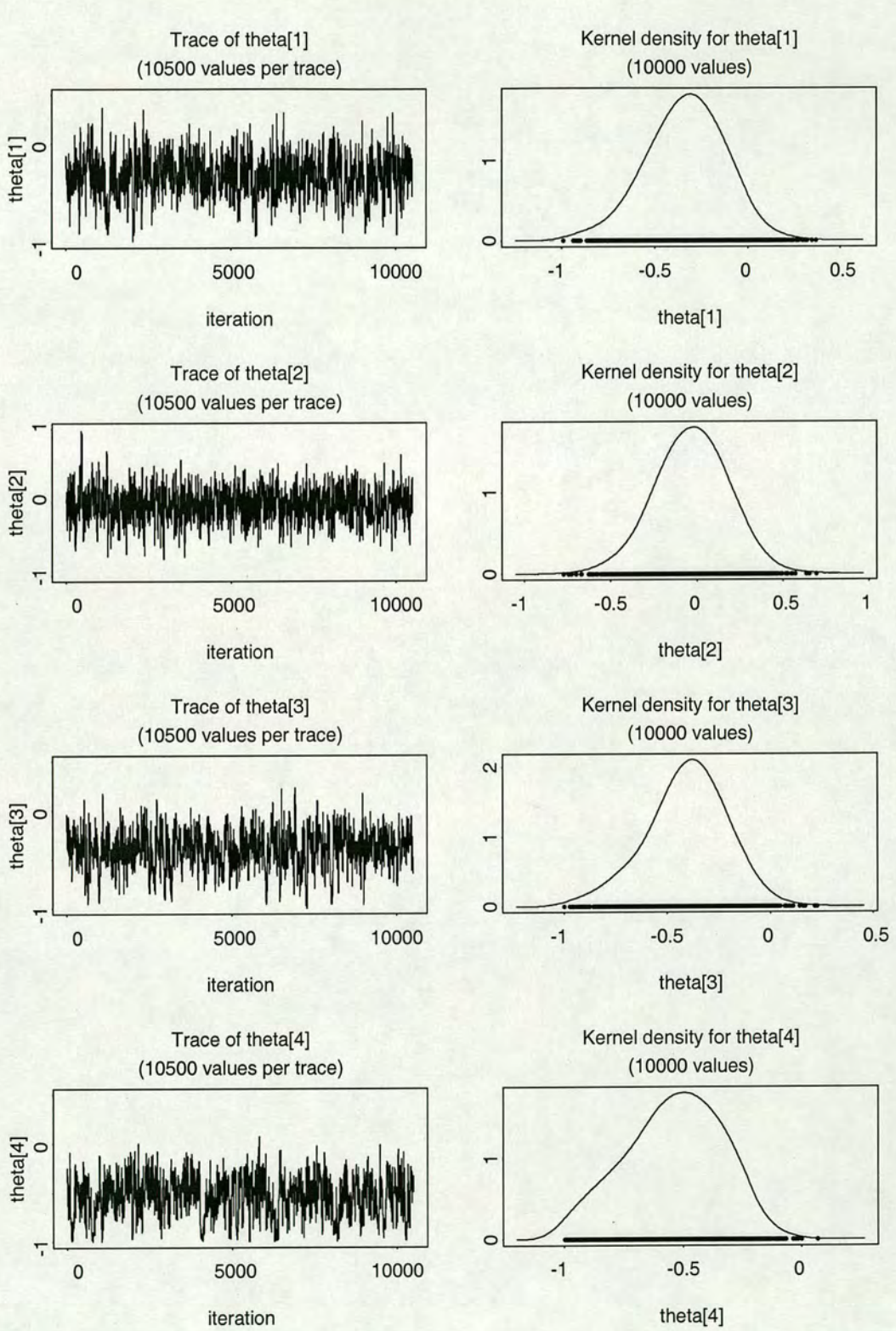

Figure B.3: *Traces of parameter estimates over the 10500 iterations and kernel density plots based on the last 10000 estimates, for each of four runs.*

```
GEWEKE CONVERGENCE DIAGNOSTIC (Z-score):
========================================

Iterations used = 501:10500
Thinning interval = 1
Sample size per chain = 10000

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

-+----------+-------------+-
 | VARIABLE |   MA3_0h    |
 | ======== |   ======    |
 |          |             |
 | theta[1] |   -0.166    |
 | theta[2] |    1.040    |
 | theta[3] |   -2.700    |
 | theta[4] |   -2.230    |
 |          |             |
-+----------+-------------+-


HEIDELBERGER AND WELCH STATIONARITY AND INTERVAL HALFWIDTH TESTS:
=================================================================

Iterations used = 501:10500
Thinning interval = 1
Sample size per chain = 10000

Precision of halfwidth test = 0.1


Chain: MA3_0h
=============

-+----------+-----------------------------------------------------+-
 |          |  Stationarity   # of iters.  # of iters.  C-vonM    |
 | VARIABLE |      test         to keep    to discard    stat.    |
 | ======== |  ============   ===========  ===========  ======    |
 |          |                                                     |
 | theta[1] |  passed         10000        0            0.1340   |
 | theta[2] |  passed         10000        0            0.0982   |
 | theta[3] |  passed         10000        0            0.2980   |
 | theta[4] |  passed         10000        0            0.4390   |
 |          |                                                     |
-+----------+-----------------------------------------------------+-
```

```
-+----------+---------------------------------+-
 |          |  Halfwidth                      |
 | VARIABLE |    test       Mean    Halfwidth |
 | ======== |  =========   ====    ========= |
 |          |                                 |
 | theta[1] |  passed      -0.3350 0.0156     |
 | theta[2] |  failed      -0.0306 0.0132     |
 | theta[3] |  passed      -0.4060 0.0152     |
 | theta[4] |  passed      -0.5390 0.0168     |
 |          |                                 |
-+----------+---------------------------------+-
```

Here, the plots of the four realisations presented in Figure B.3 display more variation then before, largely due to the short series length of 100, but the posterior kernel densities still look reasonably normal. Geweke's diagnostic provides no evidence against convergence and the Heidelberger-Welch tests are all passed, although one of the halfwidth tests is failed, indicating that the mean is not being estimated with enough precision.

It should be noted that the output presented here is obviously only a very small set of example output for illustration purposes. From inspection of all series, for all combinations of process type, parameter values, threshold level and series length, it was decided that the chains were sufficiently long for the large majority of chains to be considered as having converged.

Appendix C

Simulation results

The tables in this appendix show the detailed results from the simulation study described in Section 3.6. Figures quoted are 1000× root mean square errors (RMSE). Within a table, RMSEs can be compared down columns to see the effect of using different values of n', and across columns to compare the different methods. Figures highlighted in bold are the lowest RMSE in that column. Figures in brackets are the number out of the 100 series that the given method estimated the parameters at the boundary of the parameter space, i.e. ± 0.9999.

The methods included in the tables are:

OLS — ordinary least squares (Section 3.4.1), using either the binary *(B)* or Gaussian *(G)* autocorrelation,

Spec — spectral method (Section 3.4.5), using either the binary *(B)* or Gaussian *(G)* autocorrelation,

WLS — weighted least squares (Section 3.4.2),

GLS — generalised least squares (Section 3.4.3),

Pair — pairwise likelihood method (Section 3.4.4),

MCMC — computationally intensive method using Markov chain Monte-Carlo (Section 3.5).

Example graphs and summary tables of these results are presented within the main text in Section 3.6.2.

C.1 AR(1) processes

C.1.1 Threshold=0sd, Series length=1000

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.0	2	50.2	50.2	50.2	50.2	50.2	50.2	50.2
	4	**49.1**	**49.1**	50.4	51.3	**49.4**	49.9	**49.2**
	6	49.2	49.2	**49.6**	**50.1**	49.5	50.3	**49.2**
	10	49.2	49.2	**49.6**	50.2	49.9	50.6	**49.2**
	20	49.2	49.2	**49.6**	50.2	51.0	50.2	**49.2**
	50	49.2	49.2	**49.6**	50.2	54.4	48.1	**49.2**
	100	49.2	49.2	**49.6**	50.2	62.7	**45.3**	**49.2**
MCMC		49.3						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.3	2	**46.6**	**46.6**	**46.6**	**46.6**	**46.6**	**46.6**	**46.6**
	4	49.2	49.5	46.9	49.2	48.5	46.8	48.6
	6	49.6	50.0	47.4	47.9	48.5	47.3	48.8
	10	50.6	51.3	46.8	46.7	49.4	48.6	49.6
	20	50.6	51.3	46.8	**46.6**	49.6	50.9	49.6
	50	50.6	51.3	46.8	**46.6**	51.2	65.1	49.6
	100	50.6	51.3	46.8	**46.6**	61.1	89.8	49.6
MCMC		46.6						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.6	2	38.3	38.3	38.3	38.3	38.3	**38.3**	38.3
	4	**36.3**	**36.7**	39.9	56.2	36.2	39.0	36.1
	6	36.4	37.1	38.1	45.3	**35.8**	42.0	**35.8**
	10	39.1	40.7	36.5	40.1	36.7	49.4	36.8
	20	44.3	47.2	**35.6**	**38.2**	38.6	66.7	39.7
	50	44.9	48.1	35.7	38.3	38.8	110.4	40.0
	100	44.9	48.0	35.7	38.3	42.4	154.1	40.0
MCMC		34.4						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.9	2	24.9	24.9	24.9	**24.9**	24.9	**24.9**	24.9
	4	22.8	22.7	25.7	73.0 (35)	23.2	29.9	23.2
	6	**22.5**	**22.5**	24.5	54.9 (19)	22.7	39.3	22.7
	10	**22.5**	22.8	23.8	43.9 (11)	**22.4**	53.3	**22.4**
	20	24.5	25.4	**23.1**	30.6 (2)	23.2	89.6	23.2
	50	27.1	28.5	**23.1**	26.3	24.0	135.6	24.6
	100	28.0	29.6	23.2	**24.9**	22.7	173.7	25.1
MCMC		22.4						

C.1.2 Threshold=1sd, Series length=1000

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.0	2	74.4	75.0	74.4	75.0	75.0	75.0	75.0
	4	**73.5**	73.6	74.2	77.0	**74.0**	74.2	73.6
	6	73.7	73.6	**73.6**	75.1	74.8	72.9	**73.5**
	10	73.6	**73.5**	**73.6**	**74.9**	76.2	71.3	**73.5**
	20	73.6	**73.5**	**73.6**	75.0	79.9	67.1	**73.5**
	50	73.6	**73.5**	**73.6**	75.0	92.2	59.1	**73.5**
	100	73.6	**73.5**	**73.6**	75.0	127.0	**49.6**	**73.5**
MCMC		71.4						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.3	2	63.1	**64.6**	63.1	**64.6**	64.6	**64.6**	64.6
	4	62.6	65.8	63.4	71.2	64.5	65.2	64.8
	6	62.4	66.2	62.4	66.0	63.9	69.6	64.8
	10	61.8	65.4	61.9	64.8	62.3	73.5	64.1
	20	**61.6**	65.1	**61.7**	64.7	**61.5**	86.9	**63.9**
	50	**61.6**	65.1	**61.7**	**64.6**	66.3	114.6	**63.9**
	100	**61.6**	65.1	**61.7**	**64.6**	89.8	140.7	**63.9**
MCMC		60.1						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.6	2	45.8	50.1	45.8	**50.1**	50.1	**50.1**	50.1
	4	**42.9**	**47.7**	47.5	77.5	**46.6**	51.8	**46.6**
	6	43.3	49.4	44.7	60.0	46.7	59.3	46.7
	10	45.1	54.0	43.3	51.2	47.9	75.5	49.2
	20	47.2	59.0	**43.1**	50.3	48.1	114.7	51.6
	50	47.3	59.3	**43.1**	**50.1**	46.9	177.9	51.7
	100	47.3	59.3	**43.1**	**50.1**	52.7	232.4	51.7
MCMC		38.7						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.9	2	25.3	29.2	25.3	**29.2**	29.2	**29.2**	29.2
	4	**24.9**	**28.7**	25.8	77.4 (33)	**28.4**	40.6	28.4
	6	25.0	29.2	25.7	68.6 (25)	**28.4**	56.2	**28.2**
	10	25.9	30.6	25.1	51.0 (11)	29.0	85.6	28.8
	20	27.6	34.1	24.8	40.0 (3)	30.6	130.7	30.5
	50	29.2	38.9	**24.6**	32.9 (2)	29.9	199.6	33.1
	100	29.0	39.0	**24.6**	29.6	25.9	251.1	33.0
MCMC		21.1						

C.1.3 Threshold=0sd, Series length=100

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.0	2	**148**	**148**	148	**148**	**148**	148	**148**
	4	149	150	149	156	150	143	**148**
	6	154	155	**146**	**148**	159	142	152
	10	154	155	147	149	173	133	152
	20	154	156	147	**148**	206	119	152
	50	154	156	147	**148**	324	94	152
	100	154	156	147	**148**	466	**88**	152
MCMC		141						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.3	2	**155**	**155**	155	**155**	155	155	155
	4	158	158	156	185 (1)	155	150	156
	6	157	158	156	162	**153**	**145**	156
	10	156	157	155	158	**153**	152	**154**
	20	158	160	**153**	**155**	165	152	155
	50	158	159	**153**	**155**	209	167	**154**
	100	158	159	**153**	**155**	299	174	**154**
MCMC		145						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.6	2	146	**146**	146	146	146	**146**	146
	4	**145**	**146**	149	220 (17)	143	154	**143**
	6	**145**	147	146	180 (7)	140	168	144
	10	153	156	143	164 (3)	136	195	149
	20	155	158	**142**	**145**	120	241	149
	50	154	158	**142**	146	**98**	292	149
	100	154	157	**142**	146	115	315	149
MCMC		140						

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	GLS	*Pair*
0.9	2	94	94	94	**94**	94	**94**	94
	4	**89**	**89**	97	142 (50)	90	138	**89**
	6	93	95	91	115 (37)	92	171	91
	10	99	102	91	110 (32)	96	213	95
	20	110	115	**90**	105 (26)	98	273	102
	50	120	127	**90**	96 (9)	83	329	110
	100	121	128	**90**	**94** (1)	**58**	348	110
MCMC		106						

C.1.4 Threshold=1sd, Series length=100

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*		*WLS*	GLS	*Pair*
0.0	2	**209**	216	209	**216**		216	216	216
	4	**209**	215	208	260	(1)	**214**	193	**214**
	6	213	214	210	276	(3)	219	193	**214**
	10	213	213	**207**	377	(10)	243	164	216
	20	215	**212**	208	254	(2)	298	134	215
	50	215	**212**	208	237	(1)	434	98	215
	100	215	**212**	208	**216**		565	**81**	215
	MCMC	190							

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*		*WLS*	GLS	*Pair*
0.3	2	**218**	**221**	**218**	**221**		**221**	221	**221**
	4	231	226	224	300	(7)	222	**203**	225
	6	232	229	234	305	(4)	222	214	228
	10	229	231	234	342	(10)	229	214	227
	20	227	229	233	242	(2)	246	218	227
	50	227	229	232	285	(7)	321	218	227
	100	227	229	232	**221**		417	227	227
	MCMC	206							

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*		*WLS*	GLS	*Pair*
0.6	2	**183**	191	183	**191**		191	**191**	191
	4	185	195	183	268	(25)	191	212	194
	6	187	192	183	289	(17)	185	231	191
	10	188	192	184	215	(8)	176	276	189
	20	186	**190**	182	230	(12)	151	316	189
	50	184	**190**	**181**	212	(6)	**118**	369	**188**
	100	184	191	**181**	198	(2)	139	392	**188**
	MCMC	172							

ϕ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*		*WLS*	GLS	*Pair*
0.9	2	**151**	181	151	**181**		181	**181**	181
	4	156	176	**150**	215	(55)	177	234	164
	6	156	170	228	270	(57)	171	274	157
	10	161	169	228	194	(48)	166	335	151
	20	167	167	227	187	(25)	156	395	**143**
	50	170	160	227	178	(9)	130	468	144
	100	170	**155**	227	182	(2)	**93**	501	145
	MCMC	163							

C.2 MA(1) processes

C.2.1 Threshold=0sd, Series length=1000

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
0.0	2	**50.5**	**50.5**	**50.5**	**50.5**
	4			51.5	52.4
	6			51.5	52.4
	10			51.5	52.4
	20			51.5	52.4
	50			51.5	52.4
	100			51.5	335.2 (11)

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.3	2	**60.5**	**60.5**	60.5	60.5
	4			55.2	55.4
	6			54.7	58.4
	10			**54.2**	**54.2**
	20			**54.2**	**54.2**
	50			**54.2**	216.3 (9)
	100			**54.2**	266.9 (14)

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.6	2	**147.8** (7)	**147.9** (7)	147.8 (7)	147.9 (7)
	4			105.7 (2)	**104.3** (1)
	6			96.7 (1)	136.9 (5)
	10			95.1 (1)	139.8 (8)
	20			**94.7** (1)	140.9 (9)
	50			**94.7** (1)	240.0 (33)
	100			**94.7** (1)	244.6 (31)

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.9	2	**97.9** (91)	**175.1** (44)	97.9 (91)	175.1 (44)
	4			99.2 (98)	143.8 (38)
	6			99.8 (99)	160.1 (44)
	10			99.8 (99)	164.4 (43)
	20			99.8 (99)	152.6 (47)
	50			99.8 (99)	**138.9** (58)
	100			99.8 (99)	141.0 (49)

No MCMC results were produced here due to the large amount of computation involved.

C.2.2 Threshold=1sd, Series length=1000

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
0.0	2	**75.6**	**76.2**	**75.6**	**76.2**
	4			76.3	78.9
	6			76.2	79.3
	10			76.2	79.1
	20			76.2	79.1
	50			76.2	309.7 (9)
	100			76.2	611.1 (37)

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.3	2	**70.3**	**74.0**	70.3	74.0
	4			67.2	**72.3**
	6			**66.9**	84.7
	10			**66.9**	124.6 (2)
	20			**66.9**	220.3 (9)
	50			**66.9**	468.9 (29)
	100			**66.9**	650.8 (37)

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.6	2	**135.2** (5)	**154.5** (8)	135.2 (5)	**154.5** (8)
	4			120.8 (4)	155.1 (7)
	6			**110.1** (1)	208.6 (16)
	10			116.2 (3)	241.9 (30)
	20			115.9 (3)	254.0 (35)
	50			115.9 (3)	365.4 (47)
	100			115.9 (3)	767.8 (62)

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.9	2	**166.0** (33)	**201.0** (37)	166.0 (33)	201.0 (37)
	4			145.6 (38)	**164.3** (51)
	6			**140.0** (37)	179.3 (53)
	10			142.0 (36)	187.9 (54)
	20			142.1 (36)	184.0 (49)
	50			142.1 (36)	412.6 (54)
	100			142.1 (36)	794.6 (68)

No MCMC results were produced here due to the large amount of computation involved.

C.2.3 Threshold=0sd, Series length=100

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
0.0	2	**157**	**158**	**157**	**158**
	4			**157**	161
	6			159	178
	10			158	427 (16)
	20			158	585 (32)
	50			158	829 (67)
	100			158	1000 (100)
MCMC		185			

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.3	2	**250** (4)	**251** (4)	250 (4)	251 (4)
	4			215 (1)	**214** (2)
	6			214 (3)	242 (4)
	10			**204** (1)	401 (18)
	20			206 (1)	631 (35)
	50			206 (1)	818 (54)
	100			206 (1)	1238 (100)
MCMC		211			

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.6	2	**291** (31)	**291** (31)	291 (31)	291 (31)
	4			270 (29)	265 (28)
	6			**256** (23)	**247** (24)
	10			257 (26)	308 (32)
	20			258 (26)	577 (50)
	50			258 (26)	929 (68)
	100			258 (26)	1474 (99)
MCMC		143			

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.9	2	**187** (56)	**314** (44)	187 (56)	314 (44)
	4			154 (74)	**259** (48)
	6			**138** (71)	262 (41)
	10			**138** (73)	297 (45)
	20			**138** (73)	463 (58)
	50			**138** (73)	1085 (76)
	100			**138** (73)	1668 (100)
MCMC		247			

C.2.4 Threshold=1sd, Series length=100

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
0.0	2	**171**	**249**	**171**	**249**
	4			173	282 (1)
	6			**171**	382 (9)
	10			**171**	510 (22)
	20			**171**	632 (35)
	50			**171**	827 (63)
	100			**171**	889 (75)
MCMC		247			

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.3	2	**300** (6)	**314** (9)	**300** (6)	**314** (9)
	4			312 (6)	345 (11)
	6			305 (6)	396 (18)
	10			**300** (5)	523 (33)
	20			**300** (5)	684 (47)
	50			**300** (5)	918 (67)
	100			**300** (5)	1032 (73)
MCMC		257			

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.6	2	**327** (36)	**339** (36)	327 (36)	339 (36)
	4			**321** (33)	**333** (38)
	6			322 (33)	334 (38)
	10			**321** (36)	458 (45)
	20			**321** (36)	658 (52)
	50			**321** (36)	961 (67)
	100			**321** (36)	1176 (75)
MCMC		202			

θ	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*
−0.9	2	**321** (50)	**353** (48)	321 (50)	353 (48)
	4			280 (46)	**313** (47)
	6			**279** (41)	345 (43)
	10			280 (45)	416 (51)
	20			280 (45)	643 (50)
	50			280 (45)	978 (64)
	100			280 (45)	1275 (72)
MCMC		256			

C.3 ARMA(1,1) processes

C.3.1 Threshold=0sd, Series length=1000

$\phi = 0.0$, $\theta = -0.6$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	133		134		126		133		132		125		133	
	6	**128**		129		117		**109**		**127**		113		128	
	8	**128**		**128**		112		144		**127**		**108**		128	
	10	**128**		**128**		**109**		120		**127**		**108**		**127**	
	20	**128**		**128**		**109**		141		**127**		129		128	
	50	**128**		**128**		**109**		198		128		123		128	
	100	**128**		**128**		**109**		301		130		133		128	
θ	4	212	(17)	212	(17)	212	(17)	210	(16)	212	(17)	212	(17)	212	(17)
	6	**198**	(10)	**198**	(10)	172	(7)	174	(8)	**198**	(10)	174	(8)	**197**	(10)
	8	199	(10)	199	(10)	168	(8)	**162**	(4)	200	(10)	168	(7)	198	(10)
	10	**198**	(10)	**198**	(10)	161	(7)	172	(8)	201	(11)	**154**	(4)	198	(10)
	20	199	(10)	199	(10)	**157**	(5)	182	(8)	209	(13)	161		198	(10)
	50	199	(10)	199	(10)	158	(5)	203	(18)	237	(22)	199		198	(10)
	100	199	(10)	199	(10)	158	(5)	202	(13)	271	(32)	250		198	(10)

$\phi = 0.3$, $\theta = 0.0$

	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	*GLS*	*Pair*
ϕ	4	172	172	172	**172**	172	**172**	172
	6	**169**	**169**	**168**	175	**169**	173	**168**
	8	177	178	176	182	179	180	177
	10	178	179	174	181	179	180	178
	20	194	195	179	209	202 (1)	182	193
	50	200	201	180	191	203	207	199
	100	200	200	180	191	214	230	199
θ	4	**182**	**182**	182	**182**	**182**	182	**182**
	6	**182**	183	**176**	**182**	183	184	**182**
	8	193	194	185	190	196	193	192
	10	195	196	184	189	197	191	194
	20	218	219	189	212	234	195	216
	50	228	229	190	199	234	215	227
	100	227	228	190	199	251	240	225

$\phi = 0.3$, $\theta = -0.3$

	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	*GLS*	*Pair*
ϕ	4	**100**	**100**	100	**101**	**100**	**100**	**100**
	6	103	103	103	118	103	104	103
	8	106	107	102	124	105	114	105
	10	110	111	**99**	144	107	118	108
	20	116	118	**99**	115	113	120	112
	50	117	119	**99**	112	115	114	113
	100	117	119	**99**	122	120	124	113
θ	4	**128**	**128**	**128**	**124**	**128**	**128**	**128**
	6	134	135	142	159 (1)	133	144	133
	8	141	144	140	172	139	165 (1)	139
	10	148	151	142 (1)	194 (2)	144	158	144
	20	159	165	142 (1)	180 (2)	155	143	151
	50	161	167	142 (1)	185 (3)	165	151	152
	100	161	167	142 (1)	182 (2)	184 (1)	192	152

$\phi = 0.6, \theta = 0.3$

	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	*GLS*	*Pair*
ϕ	4	141	141	141	141	141	141	141
	6	107	107	117	133	112 (1)	115	107
	8	**103**	**103**	109	117	**103**	113	**103**
	10	108	108	**104**	**111**	108	**112**	107
	20	125	125	109	112	126	116	123
	50	133	134	113	113	134 (1)	121	132
	100	133	134	116	114	131 (1)	119	132
θ	4	170	170	170	170	170	170	170
	6	134	134	138	151	147	142	133
	8	**126**	**127**	131	138	**127**	**136**	**125**
	10	135	137	**125**	**131**	137	137	133
	20	163	165	134	135	168	142	160
	50	178	180	142	137	186	154	175
	100	178	180	147	138	185	157	175

$\phi = 0.6, \theta = -0.3$

	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	*GLS*	*Pair*
ϕ	4	**48**	**48**	**48**	**40**	**48**	**48**	**48**
	6	51	52	51	85	50	52	50
	8	57	59	50	93	55	67	55
	10	63	65	49	118	59	73	60
	20	76	81	**48**	90	68	78	70
	50	79	85	**48**	80	69	80	72
	100	79	85	**48**	72	71	84	72
θ	4	**96**	**96**	96	**72**	**96**	**96**	**96**
	6	119	138	97	132	109	111	109
	8	170 (1)	221 (5)	98	128	138	141	141
	10	217 (4)	258 (7)	**92**	174 (1)	165 (1)	145	168 (1)
	20	299 (10)	354 (15)	96	180 (1)	219 (2)	159	224 (2)
	50	308 (10)	367 (15)	96	193 (4)	221 (2)	225	229 (2)
	100	308 (10)	367 (15)	96	190 (3)	217 (2)	315	229 (2)

$\phi = 0.9, \theta = 0.6$

	n'	*OLS (B)*	*OLS (G)*	*Spec (B)*	*Spec (G)*	*WLS*	*GLS*	*Pair*
ϕ	4	82 (12)	82 (14)	82 (14)	82 (15)	82 (15)	82 (17)	82 (10)
	6	56 (1)	56 (1)	58 (1)	57 (1)	69 (28)	58 (7)	56 (1)
	8	46	46	47	48	57 (17)	48 (2)	45
	10	36	36	38	39	49 (12)	**37**	36
	20	**35**	**35**	34	36	**35**	41	**34**
	50	38	39	**32**	**34**	37	43	37
	100	39	40	**32**	**34**	36	45	38
θ	4	204	204	204	205	205	207	203
	6	132	132	132	132	223	153	132
	8	97	98	100	100	178	113	97
	10	73	**73**	75	75	151	76	73
	20	**71**	**73**	67	72	**72**	**72**	**68**
	50	87	92	**62**	**66**	84	**72**	81
	100	93	99	63	**66**	91	83	85

No MCMC results were produced here due to the large amount of computation involved.

C.3.2 Threshold=1sd, Series length=1000

$\phi = 0.0, \theta = -0.6$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	**127**		**145**		**116**		175		**139**		**122**		**141**	
	6	134		148		122		**127**		143		134		146	
	8	134		147		123		233		142		138		144	
	10	134		148		121		**127**		142		146		145	
	20	135		149		120		151		142		150		146	
	50	135		149		120		265		140		172		146	
	100	135		149		120		438		151		172		146	
θ	4	**307**	(49)	**256**	(30)	**307**	(49)	250	(25)	**256**	(30)	256	(30)	**256**	(30)
	6	312	(50)	264	(30)	317	(55)	253	(24)	264	(30)	**252**	(25)	263	(30)
	8	313	(51)	258	(27)	309	(51)	**215**	(11)	262	(29)	254	(22)	257	(27)
	10	314	(51)	261	(28)	310	(50)	269	(30)	265	(29)	261	(17)	259	(28)
	20	314	(51)	261	(27)	312	(52)	272	(25)	278	(34)	259	(3)	259	(27)
	50	314	(51)	261	(27)	312	(52)	284	(37)	305	(46)	345		259	(26)
	100	314	(51)	261	(27)	312	(52)	290	(36)	352	(64)	405		258	(26)

$\phi = 0.3, \theta = 0.0$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	269		272	(1)	269	(1)	272	(1)	272	(1)	272	(1)	272	
	6	271		270		**265**		270		280	(3)	276		269	
	8	**260**	(1)	**261**	(1)	268	(1)	279	(1)	**262**	(1)	268	(1)	**261**	(1)
	10	269		269		270		291		272	(1)	278		268	
	20	275		274		274		**230**		275		268		273	
	50	275		273		273		248		273		264		272	
	100	275		273		273		258		278		**245**		272	
θ	4	310		308		310		308		308		308		308	
	6	307	(1)	302	(1)	299	(1)	**297**	(1)	325	(1)	310	(1)	301	(1)
	8	**294**		**291**		299		307	(1)	293		297		**290**	
	10	296	(1)	296	(2)	**298**	(1)	316	(1)	**305**	(2)	307	(2)	296	(2)
	20	307	(1)	304	(2)	301	(2)	352	(6)	311	(2)	294	(1)	303	(2)
	50	307	(2)	302	(2)	301	(1)	346	(4)	319	(2)	295		301	(2)
	100	307	(2)	302	(2)	300		345	(4)	350	(4)	**277**		301	(2)

$\phi = 0.3, \theta = -0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	**111**		**115**		**110**		127		**114**		**112**		**114**	
	6	119		124		116		141		122		123		123	
	8	119		126		117		155		123		130		124	
	10	123		132		115		166		127		150		128	
	20	131		145		117		123		141		130		138	
	50	132		145		118		**121**		143		133		138	
	100	132		145		118		188		155		127		138	
θ	4	**226**	(3)	**195**	(2)	**226**	(3)	**150**		**195**	(2)	195	(2)	**195**	(2)
	6	237	(3)	212	(2)	247	(7)	231	(4)	209	(2)	226	(4)	209	(2)
	8	242	(3)	218	(2)	242	(5)	210		213	(2)	212	(2)	212	(2)
	10	246	(3)	227	(2)	242	(5)	314	(14)	219	(2)	232	(2)	217	(2)
	20	257	(3)	247	(2)	245	(5)	304	(9)	245	(2)	**194**		233	(2)
	50	258	(3)	247	(2)	245	(5)	310	(13)	259	(3)	259		233	(2)
	100	258	(3)	247	(2)	245	(5)	309	(11)	274	(5)	304		233	(2)

$\phi = 0.6, \theta = 0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	217	(3)	220	(3)	217	(3)	220	(3)	220	(3)	220	(4)	220	(3)
	6	165	(1)	168	(1)	199	(1)	232	(1)	178	(5)	193	(3)	168	(1)
	8	144		147		190		229		160	(3)	**164**		147	
	10	141		145		180		219		151	(2)	167	(1)	144	
	20	136		142		178		202		147	(1)	167		141	
	50	**134**		**139**		**177**		**174**		133		175		**138**	
	100	**134**		**139**		**177**		205		**127**		173		**138**	
θ	4	302		297		302		297		297		298		297	
	6	235		230		263		289		252		263		231	
	8	192		189		235		271		219		211		188	
	10	189		188		222		**258**		206		218		186	
	20	187		190		217		289	(2)	208		**205**		186	
	50	**184**		185		**216**		289	(2)	191		206		**181**	
	100	**184**		**184**		**216**		284	(1)	**187**		208		**181**	

$\phi = 0.6, \theta = -0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	**62**		**69**		61		**53**		**69**		**68**		**69**	
	6	**62**		71		65		128		**69**		82		**69**	
	8	63		73		63		126		70		99		70	
	10	63		76		62		272		71		105		72	
	20	66		85		**59**		124		74		114		76	
	50	66		85		**59**		118		72		112		77	
	100	66		84		**59**		102	(1)	76		112		76	
θ	4	**235**	(5)	**201**	(2)	235	(5)	**110**		**201**	(2)	**201**	(2)	**201**	(2)
	6	245	(4)	232	(3)	248	(4)	289	(10)	211	(3)	233	(2)	212	(3)
	8	276	(8)	271	(8)	238	(5)	151		232	(4)	251	(3)	232	(4)
	10	289	(9)	317	(12)	232	(5)	339	(18)	249	(6)	234	(1)	250	(6)
	20	315	(10)	379	(17)	232	(5)	253	(1)	267	(6)	258		271	(6)
	50	316	(10)	382	(18)	**231**	(5)	304	(10)	263	(6)	352		271	(6)
	100	316	(10)	381	(18)	**231**	(5)	301	(10)	255	(6)	418		271	(6)

$\phi = 0.9, \theta = 0.6$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	105	(23)	101	(26)	105	(26)	101	(26)	101	(26)	101	(28)	101	(23)
	6	76	(9)	72	(9)	80	(9)	77	(9)	83	(37)	76	(18)	72	(9)
	8	61	(2)	57	(2)	65	(2)	73	(2)	71	(32)	62	(10)	56	(2)
	10	53	(1)	49	(1)	57	(1)	58	(1)	63	(23)	56	(5)	49	(1)
	20	**49**		**46**		51		51		49	(4)	**51**		**46**	
	50	52		51		**49**		48		47		52		49	
	100	53		53		**49**		**47**		**45**	(2)	78		51	
θ	4	250		249		250		250		250		253		249	
	6	179		178		188		186		256		204		178	
	8	131		130		139		146		234		161		129	
	10	109		109		113		115		202		130		108	
	20	**81**		**83**		86		94		113		**86**		**82**	
	50	85		98		79		86		**96**		87		89	
	100	91		110		**78**		**84**		119		120		97	

No MCMC results were produced here due to the large amount of computation involved.

C.3.3 Threshold=0sd, Series length=100

$\phi = 0.0, \theta = -0.6$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	328		329		315		384		324		**304**		328	
	6	341		341		291		**277**		360	(4)	325	(3)	339	
	8	319		320		311		368		314		334	(2)	317	
	10	**310**		**312**		313		368		**307**		318	(2)	**308**	
	20	316		317		310		319		350	(4)	344	(1)	314	
	50	316		316		**306**		391		388	(2)	353		314	
	100	314		315		**306**		501		582	(14)	355		315	
θ	4	380	(48)	380	(48)	380	(48)	370	(30)	**380**	(48)	**380**	(48)	380	(48)
	6	415	(31)	416	(31)	**340**	(33)	338	(33)	471	(34)	443	(22)	412	(31)
	8	377	(37)	380	(36)	377	(30)	**322**	(11)	387	(44)	498	(18)	373	(37)
	10	**369**	(34)	**374**	(36)	377	(34)	340	(30)	382	(44)	456	(17)	**361**	(32)
	20	388	(35)	394	(37)	367	(35)	328	(24)	499	(55)	487	(5)	382	(36)
	50	389	(35)	393	(37)	367	(35)	341	(29)	527	(59)	505		383	(35)
	100	385	(35)	389	(37)	367	(35)	340	(25)	783	(40)	528	(1)	384	(35)

MCMC ϕ - 173 θ - 356

$\phi = 0.3, \theta = 0.0$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	504	(10)	504	(10)	499	(9)	495	(10)	503	(10)	493	(10)	503	(10)
	6	432	(6)	432	(6)	453	(8)	446	(8)	433	(11)	416	(10)	433	(6)
	8	425	(2)	425	(2)	401	(5)	391	(4)	445	(12)	**388**	(5)	424	(2)
	10	423	(3)	423	(3)	405	(1)	400	(1)	436	(10)	436	(5)	422	(3)
	20	399	(1)	399	(1)	379	(1)	332	(1)	408	(9)	403	(3)	398	(1)
	50	392		391		**371**		**311**		**402**	(7)	455		392	
	100	**388**		**388**		**371**		325		571	(40)	448		**386**	
θ	4	609	(16)	609	(16)	609	(16)	606	(16)	609	(17)	610	(17)	609	(16)
	6	541	(5)	542	(5)	539	(8)	535	(8)	**549**	(5)	523	(5)	541	(5)
	8	515	(10)	516	(10)	481	(5)	467	(5)	573	(10)	482	(5)	514	(9)
	10	501	(6)	502	(6)	460	(3)	474	(6)	559	(9)	493	(4)	499	(6)
	20	482	(6)	483	(6)	440	(4)	**427**	(5)	566	(13)	454	(2)	483	(6)
	50	473	(6)	473	(6)	432	(4)	451	(8)	581	(18)	441	(3)	472	(6)
	100	**466**	(6)	**467**	(6)	**431**	(4)	472	(13)	762	(13)	**438**		**469**	(6)

MCMC ϕ - 323 θ - 189

$\phi = 0.3, \theta = -0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	310	(1)	311	(1)	**294**	(1)	304	(1)	307	(1)	290	(1)	308	(1)
	6	308		309		297		303		312	(4)	298	(4)	306	
	8	**299**		**301**		318		310		307	(3)	284	(1)	**297**	
	10	303	(1)	304	(1)	318		347		318	(6)	**282**	(3)	301	(1)
	20	310	(1)	312	(1)	302		283	(1)	314	(5)	323		306	(1)
	50	302		304		304		**253**		**298**	(3)	344		302	
	100	300		**301**		304		310		362	(6)	319		299	
θ	4	**455**	(24)	**456**	(24)	455	(24)	358	(11)	**456**	(24)	456	(24)	**456**	(24)
	6	489	(21)	491	(22)	410	(16)	400	(16)	509	(23)	453	(12)	486	(21)
	8	463	(24)	467	(24)	426	(21)	**338**	(5)	503	(27)	**411**	(9)	461	(24)
	10	489	(22)	496	(24)	423	(21)	467	(29)	545	(26)	434	(9)	484	(22)
	20	505	(22)	516	(25)	**409**	(15)	416	(11)	566	(34)	419	(2)	496	(22)
	50	491	(22)	500	(25)	411	(17)	433	(19)	565	(41)	425		487	(23)
	100	485	(22)	495	(25)	411	(17)	449	(22)	599	(27)	420		481	(23)

MCMC ϕ - 224 θ - 173

$\phi = 0.6, \theta = 0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	471	(14)	471	(15)	471	(14)	470	(15)	471	(15)	468	(22)	471	(14)
	6	466	(6)	466	(6)	450	(8)	442	(8)	477	(22)	489	(14)	466	(6)
	8	448	(4)	448	(4)	437	(5)	411	(4)	458	(17)	**428**	(8)	447	(4)
	10	440	(2)	440	(2)	420	(1)	418	(1)	452	(16)	**428**	(4)	439	(2)
	20	444		444		420		384		437	(9)	501	(3)	444	
	50	438		439		**417**		**375**		**390**	(8)	513	(1)	438	
	100	**437**		**437**		**417**		378		403	(28)	499	(1)	**436**	
θ	4	582	(5)	582	(5)	583	(5)	583	(5)	583	(5)	586	(5)	582	(5)
	6	563	(6)	565	(6)	566	(7)	578	(8)	603	(6)	584	(5)	564	(6)
	8	531	(4)	535	(4)	528	(4)	535	(5)	576	(4)	525	(3)	529	(4)
	10	524	(4)	531	(5)	486	(2)	511	(4)	578	(4)	502	(2)	520	(4)
	20	531	(6)	536	(6)	489	(3)	**494**	(4)	583	(6)	504	(3)	523	(5)
	50	525	(6)	529	(6)	486	(3)	509	(4)	565	(8)	484		516	(5)
	100	**521**	(6)	**525**	(6)	**485**	(2)	514	(5)	**541**	(4)	**457**		**513**	(5)

MCMC ϕ - 431 θ - 354

$\phi = 0.6, \theta = -0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	**192**	(1)	**193**	(1)	186	(1)	**143**	(1)	190	(1)	**191**	(2)	**190**	(1)
	6	205		208		191		257		203	(2)	200	(1)	201	
	8	216		219		**179**		170		208	(1)	222		209	
	10	218		221		188		365		204		220		211	
	20	222		226		183		221	(2)	201	(3)	218		214	
	50	220		224		181		188		166	(1)	229		213	
	100	218		221		181		176	(2)	**160**	(2)	230		211	
θ	4	**448**	(21)	**449**	(21)	448	(21)	286		**449**	(21)	446	(20)	**449**	(21)
	6	490	(30)	499	(31)	434	(20)	386	(13)	498	(28)	414	(13)	479	(28)
	8	494	(31)	516	(38)	414	(21)	**260**		495	(31)	398	(7)	477	(30)
	10	509	(38)	522	(40)	419	(21)	463	(29)	498	(36)	**375**	(4)	484	(31)
	20	534	(42)	546	(44)	415	(20)	405	(5)	553	(40)	393		511	(37)
	50	532	(43)	546	(45)	414	(21)	474	(29)	512	(37)	442		512	(37)
	100	523	(43)	537	(45)	**413**	(21)	461	(22)	475	(19)	464		507	(38)

MCMC ϕ - 187 θ - 202

$\phi = 0.9, \theta = 0.6$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	367	(30)	367	(30)	367	(30)	366	(30)	367	(30)	367	(46)	367	(29)
	6	**299**	(20)	**300**	(20)	401	(22)	462	(21)	298	(45)	**348**	(35)	**300**	(20)
	8	302	(16)	303	(16)	338	(20)	381	(16)	322	(54)	411	(32)	301	(15)
	10	322	(9)	323	(9)	337	(12)	398	(10)	322	(47)	403	(27)	322	(9)
	20	321	(1)	322	(1)	**324**	(1)	373		321	(39)	459	(10)	319	(1)
	50	320		322		**324**		325		270	(14)	438		319	
	100	321		323		**324**		**315**		**236**	(16)	404		319	
θ	4	500		501		502		495		501		507		500	
	6	405		405		539	(3)	603	(6)	434		460	(1)	406	
	8	**396**		**402**		472	(2)	502	(1)	462		487		**394**	
	10	413		429	(1)	444	(1)	**480**		460		479		407	
	20	417		429	(1)	417	(1)	487	(2)	475	(1)	445		397	
	50	429		458	(1)	414	(1)	491	(3)	422	(1)	391		401	
	100	442		539	(4)	**413**	(1)	489	(3)	**391**	(1)	**359**		406	

MCMC ϕ - 434 θ - 422

C.3.4 Threshold=1sd, Series length=100

$\phi = 0.0, \theta = -0.6$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	429		417		416		463		411		**387**	(1)	517	(10)
	6	416	(2)	424	(1)	405	(1)	437	(1)	416	(1)	436	(5)	488	(6)
	8	414	(1)	437	(3)	397	(1)	466	(1)	432	(4)	422	(4)	466	(2)
	10	399	(1)	408	(1)	392	(1)	**416**	(1)	414	(6)	406	(4)	437	(1)
	20	385		**391**	(1)	381		436		**406**	(3)	426	(2)	429	(1)
	50	391		**391**	(1)	381		467	(1)	556	(16)	430	(2)	**427**	(1)
	100	**383**		**391**		**380**		557		775	(46)	426	(3)	**427**	
θ	4	521	(68)	510	(54)	521	(68)	491	(33)	**512**	(54)	**533**	(51)	708	(55)
	6	546	(56)	558	(48)	520	(52)	546	(49)	555	(50)	634	(25)	663	(46)
	8	545	(56)	579	(52)	516	(53)	479	(15)	591	(53)	624	(21)	606	(51)
	10	526	(54)	540	(49)	507	(52)	504	(41)	588	(57)	604	(18)	564	(47)
	20	489	(52)	503	(49)	487	(50)	**463**	(22)	548	(62)	600	(9)	544	(47)
	50	486	(55)	498	(48)	479	(51)	464	(29)	791	(51)	594	(2)	**536**	(47)
	100	**477**	(54)	**496**	(48)	**478**	(52)	466	(25)	1147	(31)	573	(1)	537	(46)

MCMC ϕ - 166 θ - 442

$\phi = 0.3, \theta = 0.0$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	646	(8)	624	(7)	632	(8)	623	(10)	621	(10)	591	(19)	646	(13)
	6	576	(7)	567	(8)	594	(7)	569	(10)	573	(17)	548	(14)	577	(13)
	8	544	(4)	**519**	(5)	545	(5)	477	(4)	546	(14)	**484**	(10)	**528**	(6)
	10	565	(3)	543	(4)	551	(1)	498	(4)	558	(16)	531	(9)	556	(6)
	20	546	(2)	520	(3)	**541**	(2)	448	(2)	**514**	(9)	491	(3)	**528**	(4)
	50	547		526		**541**		**426**		558	(24)	507	(3)	**528**	(2)
	100	**539**		529		553		477		717	(67)	522		**528**	(2)
θ	4	825	(39)	778	(30)	824	(39)	772	(27)	779	(30)	783	(30)	820	(31)
	6	735	(29)	725	(24)	734	(30)	729	(28)	751	(25)	662	(15)	741	(23)
	8	718	(31)	678	(23)	704	(28)	613	(11)	735	(28)	579	(10)	680	(21)
	10	726	(30)	681	(23)	697	(26)	649	(20)	744	(29)	602	(9)	687	(22)
	20	694	(26)	**666**	(26)	682	(26)	**603**	(14)	**714**	(31)	510	(6)	658	(23)
	50	691	(29)	672	(26)	**676**	(27)	630	(21)	775	(32)	490	(4)	**654**	(22)
	100	**685**	(28)	679	(26)	682	(28)	639	(18)	886	(10)	**472**	(1)	657	(22)

MCMC ϕ - 279 θ - 152

$\phi = 0.3, \theta = -0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	434	(1)	391	(1)	401	(1)	400	(1)	380	(1)	**347**	(2)	433	(9)
	6	408		387	(1)	386	(1)	395	(2)	385	(5)	358	(7)	406	(4)
	8	400		387	(1)	372		348	(1)	383	(6)	357	(6)	389	(1)
	10	397		394	(2)	366		379	(1)	382	(8)	387	(8)	390	(2)
	20	393	(1)	386	(1)	367	(1)	360	(4)	**378**	(11)	378	(2)	385	(3)
	50	387		**383**	(3)	365		**338**	(2)	391	(15)	384	(3)	382	(2)
	100	**386**		388	(2)	**362**		382	(1)	556	(43)	399	(5)	**381**	(2)
θ	4	580	(47)	**539**	(37)	580	(47)	**443**	(19)	**540**	(37)	**522**	(33)	645	(37)
	6	562	(40)	572	(36)	556	(41)	521	(31)	601	(35)	560	(17)	612	(34)
	8	560	(40)	575	(34)	552	(40)	427	(11)	611	(33)	529	(11)	574	(33)
	10	568	(37)	585	(35)	546	(38)	524	(34)	625	(34)	559	(12)	581	(33)
	20	555	(38)	561	(34)	539	(37)	458	(13)	662	(38)	523		561	(32)
	50	542	(38)	571	(33)	534	(38)	478	(21)	707	(37)	522		554	(32)
	100	**540**	(38)	574	(34)	**530**	(37)	476	(14)	932	(20)	523		**552**	(32)

MCMC ϕ - 225 θ - 205

$\phi = 0.6, \theta = 0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	727	(23)	701	(24)	742	(25)	718	(24)	699	(25)	648	(37)	701	(26)
	6	616	(13)	606	(15)	679	(15)	632	(17)	620	(30)	588	(29)	603	(17)
	8	622	(9)	630	(11)	642	(10)	590	(12)	653	(34)	**574**	(24)	626	(14)
	10	631	(6)	639	(8)	641	(4)	606	(8)	639	(28)	669	(14)	626	(12)
	20	623	(3)	618	(4)	637	(3)	573	(6)	592	(23)	592	(5)	610	(5)
	50	609		**596**		634		**528**	(2)	**560**	(33)	613	(3)	582	(3)
	100	**605**		**596**	(2)	**631**		545	(1)	575	(74)	644	(6)	**581**	(3)
θ	4	834	(22)	801	(15)	849	(20)	817	(17)	802	(15)	783	(15)	810	(15)
	6	706	(15)	697	(14)	790	(20)	791	(21)	735	(14)	670	(8)	703	(14)
	8	676	(13)	688	(12)	705	(14)	711	(7)	754	(13)	619	(4)	687	(12)
	10	687	(14)	700	(14)	695	(16)	746	(18)	734	(12)	670	(7)	681	(12)
	20	665	(15)	677	(15)	679	(12)	719	(12)	704	(13)	570	(5)	658	(13)
	50	642	(12)	**657**	(14)	669	(12)	715	(13)	**702**	(14)	**536**	(2)	633	(12)
	100	**639**	(13)	**657**	(14)	**667**	(12)	**706**	(7)	708	(3)	549		**629**	(12)

MCMC ϕ - 489 θ - 345

$\phi = 0.6, \theta = -0.3$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	293	(1)	290	(1)	252	(1)	**250**	(1)	275	(1)	**259**	(6)	289	(6)
	6	282	(1)	282	(1)	266	(1)	375	(1)	268	(3)	269	(3)	288	(9)
	8	277		**281**	(1)	**243**		263	(4)	261	(2)	299	(2)	**284**	(9)
	10	280		284	(1)	247		359	(2)	263	(6)	302	(3)	287	(9)
	20	278		302	(1)	249	(1)	275	(4)	246	(7)	294	(1)	287	(9)
	50	**273**		289	(1)	244		263	(2)	**231**	(16)	312	(2)	**284**	(6)
	100	**273**		314	(2)	262		277	(1)	245	(23)	351	(2)	286	(6)
θ	4	599	(46)	560	(35)	599	(46)	395	(5)	561	(35)	556	(28)	**627**	(35)
	6	567	(42)	**543**	(37)	566	(46)	458	(24)	549	(35)	483	(15)	639	(34)
	8	**555**	(47)	551	(42)	534	(39)	**341**	(5)	**544**	(39)	**458**	(13)	638	(36)
	10	563	(51)	574	(44)	550	(45)	468	(27)	592	(40)	463	(5)	653	(39)
	20	575	(56)	610	(50)	546	(44)	432	(8)	620	(42)	495		670	(42)
	50	567	(57)	637	(50)	534	(44)	436	(16)	700	(38)	587		650	(43)
	100	567	(57)	661	(50)	**533**	(44)	454	(17)	736	(21)	570		663	(42)

MCMC ϕ - 240 θ - 179

$\phi = 0.9, \theta = 0.6$

	n'	*OLS (B)*		*OLS (G)*		*Spec (B)*		*Spec (G)*		*WLS*		*GLS*		*Pair*	
ϕ	4	677	(23)	678	(25)	723	(25)	715	(26)	676	(26)	668	(44)	662	(30)
	6	**592**	(13)	588	(14)	677	(17)	724	(19)	595	(40)	549	(45)	578	(18)
	8	610	(13)	569	(14)	652	(11)	609	(12)	561	(42)	580	(28)	569	(15)
	10	611	(6)	561	(10)	650	(9)	664	(12)	538	(45)	574	(29)	559	(12)
	20	599	(4)	548	(6)	653	(3)	593	(8)	525	(41)	571	(19)	546	(12)
	50	597		**542**	(3)	**647**		578	(1)	**491**	(35)	513	(6)	**540**	(7)
	100	598		546	(3)	**647**		**571**		529	(49)	**512**	(5)	545	(2)
θ	4	779	(18)	768	(10)	824	(16)	816	(14)	769	(10)	773	(10)	761	(10)
	6	666	(12)	665	(9)	787	(17)	928	(20)	691	(9)	635	(4)	659	(9)
	8	661	(7)	630	(5)	717	(13)	754	(7)	649	(5)	634	(4)	626	(5)
	10	654	(10)	**615**	(6)	726	(13)	818	(14)	626	(5)	637	(5)	609	(6)
	20	**648**	(9)	633	(6)	697	(14)	**748**	(6)	607	(5)	534		590	(5)
	50	668	(10)	690	(9)	**692**	(14)	787	(13)	**581**	(6)	435	(1)	**586**	(6)
	100	673	(11)	698	(9)	**692**	(14)	784	(9)	585	(4)	**422**		588	(6)

MCMC ϕ - 554 θ - 511

Bibliography

Abramowitz, M. and Stegun, I. A., editors (1965). *Handbook of mathematical functions: with formulas, graphs and mathematical tables.* Dover Publications, New York.

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B*, 53:111–142.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19:716–723.

Alados, C. L., Escos, J. M., and Emlen, J. M. (1996). Fractal structure of sequential behaviour patterns: an indicator of stress. *Animal Behaviour*, 51:437–443.

Albert, P. S. (1991). A two-state Markov model for a time series of epileptic seizure counts. *Biometrics*, 47:1371–1381.

Allcroft, D. J. and Glasbey, C. A. (2000). Estimation of latent Gaussian ARMA models for categorical behaviour data. In V. Núñez-Antón and E. Ferreira, editors, *Proceedings of the 15th International Workshop on Statistical Modelling*, pages 294–299. The University of the Basque Country, Bilbao.

Allcroft, D. J. and Glasbey, C. A. (2001). A spectral estimator of ARMA parameters from thresholded data. *Statistics and Computing.* (in press).

Allcroft, D. J., Glasbey, C. A., Kyriazakis, I., and Tolkamp, B. J. (1999). What is the critical timescale of a cow's feeding behaviour? in Proceedings of the Thirty-first Meeting of the Agricultural Research Modellers' Group. *Journal of Agricultural Science*, 133:344–345.

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66:59–65.

Atkinson, A. C. (1970). A method for discriminating between models (with discussion). *Journal of the Royal Statistical Society, Series B*, 32:323–353.

Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford University Press, Oxford.

Bak, P. (1997). *How Nature Works: the Science of Self-Organized Criticality.* Oxford University Press, Oxford.

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis.* Arnold, London, second edition.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society, Series B*, 8:27–41.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73:360–363.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

Berdoy, M. (1993). Defining bouts of behaviour: a three-process model. *Animal Behaviour*, 46:387–396.

Best, N. G., Cowles, M. K., and Vines, S. K. (1995). *CODA: Convergence Diagnostics and Output Analysis Software for Gibbs sampling output, Version 0.30.* MRC Biostatistics Unit, Cambridge, UK.

Best, N. G., Cowles, M. K., and Vines, S. K. (1997). *CODA: Convergence Diagnostics and Output Analysis Software for Gibbs sampling output, Version 0.40 (Addendum to Manual).* MRC Biostatistics Unit, Cambridge, UK.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control.* Prentice Hall, Englewood Cliffs, New Jersey, third edition.

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory.* Holt, Rinehart and Winston, New York.

Brooks, S. P. and Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8:319–335.

Burden, R. L. and Faires, J. D. (1985). *Numerical Analysis.* Prindle, Weber & Schmidt, Boston, third edition.

Chandler, R. E. (1996). A note on analytical solutions to the Whittle likelihood equation. Technical Report 173, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT.

Chatfield, C. (1979). Inverse autocorrelations. *Journal of the Royal Statistical Society, Series A*, 142:363–377.

Chatfield, C. (1996). *The Analysis of Time Series: An Introduction.* Chapman and Hall, London, fifth edition.

Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics*, 14:277–293.

Cole, B. J. and Cheshire, D. (1996). Mobile cellular automata models of ant behavior: movement activity of *leptothorax allardycei*. *The American Naturalist*, 148:1–15.

Collier, G., Johnson, D. F., and Mitchell, C. (1999). The relation between meal size and the time between meals: effects of cage complexity and food cost. *Physiology and Behavior*, 67:339–346.

Coursol, J. and Dacunha-Castelle, D. (1983). Remarks on the approximation of the likelihood function of a stationary Gaussian process. *Theory of Probability and its Applications*, 27:162–167.

Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the fourth Berkeley Symposium*, volume 1, pages 105–123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, 24:406–424.

Cox, D. R. (1970). *Renewal Theory*. Methuen & Co. Ltd., UK.

Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman & Hall, London.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Methuen, London.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Chapman & Hall, London, second edition.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 17:563–586.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

De Castro, J. M. (1975). Meal pattern correlations: facts and artifacts. *Physiology and Behavior*, 15:13–15.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Ermentrout, G. B. and Edelstein-Keshet, L. (1993). Cellular automata approaches to biological modeling. *Journal of Theoretical Biology*, 160:97–133.

Fagen, R. M. and Young, D. Y. (1978). Temporal patterns of behaviors: durations, intervals, latencies, and sequences. In Colgan, P. W., editor, *Quantitative Ethology*. Wiley, New York.

Ferriere, R., Cazelles, B., Cezilly, F., and Desportes, J.-P. (1996). Predictability and chaos in bird vigilant behaviour. *Animal Behaviour*, 52:457–472.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data.* The MIT Press, Cambridge, Massachusetts, second edition.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall, London.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 599–607. Oxford University Press, Oxford.

Glasbey, C. A. and Martin, R. J. (1986). Exploratory and confirmatory plots of single-channel records. *Journal of Neuroscience Methods*, 16:239–249.

Glasbey, C. A. and McGechan, M. B. (1986). The assessment of combining work-days criteria and forecasting models. *Journal of Agricultural Engineering Research*, 33:23–31.

Glasbey, C. A. and Nevison, I. M. (1997). Rainfall modelling using a latent Gaussian variable. In Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen, E., Warren, W. G., and Wolfinger, R. D., editors, *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, number 122 in Lecture Notes in Statistics, pages 233–242. Springer, New York.

Glasbey, C. A., Nevison, I. M., and Hunter, A. G. M. (1998). Parameter estimators for Gaussian models with censored time series and spatio-temporal data. In Payne, R. and Green, P., editors, *COMPSTAT98 Proceedings in Computational Statistics*, pages 323–328, Heidelberg. Physica-Verlag.

Green, P. J. (1981). Peeling bivariate data. In Barnett, V., editor, *Interpreting Multivariate Data*, pages 3–19. John Wiley, Chichester.

Haccou, P. and Meelis, E. (1994). *Statistical Analysis of Behavioural Data: An Approach Based on Time-structured Models.* Oxford University Press, Oxford.

Harley, C. B. (1981). Learning the evolutionarily stable strategy. *Journal of Theoretical Biology*, 89:611–633.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93:1099–1111.

Hinde, J. P. (1992). Choosing between non-nested models: a simulation approach. In *Proceedings of GLIM92 Conference*, Munich.

Hjort, N. L. and Omre, H. (1994). Topics in spatial statistics. *Scandinavian Journal of Statistics*, 21:289–357.

Houston, A. I. and Sumida, B. H. (1987). Learning rules, matching and frequency dependence. *Journal of Theoretical Biology*, 126:289–308.

Johnson, N. L. and Kotz, S. (1972). *Continuous Multivariate Distributions.* Wiley, New York.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kedem, B. (1980). *Binary Time Series.* Lecture notes in pure and applied mathematics; v.52. Dekker, New York.

Keenan, D. (1982). A time series analysis of binary data. *Journal of the American Statistical Association*, 77:816–821.

Kendall, M. G., Stuart, A., and Ord, J. K. (1983). *The Advanced Theory of Statistics, Volume 3: Design and Analysis, and Time-Series.* Griffin, High Wycombe, fourth edition.

Kent, J. T. (1986). The underlying structure of nonnested hypothesis tests. *Biometrika*, 73:333–343.

Kyriazakis, I. (1997). The nutritional choices of farm animals: to eat or what to eat? In Forbes, J. M., Lawrence, T. L. J., Rodaway, R. G., and Varley, M. A., editors, *Animal Choices*, pages 55–65. Occasional Publication No. 20 — British Society of Animal Science.

Langton, S. D., Collett, D., and Sibly, R. M. (1995). Splitting behaviour into bouts; a maximum likelihood approach. *Behaviour*, 132:781–799.

Lawes Agricultural Trust (1993). *Genstat 5 Release 3 Reference Manual.* Oxford University Press, Oxford.

Le, N. D., Leroux, B. G., and Puterman, M. L. (1992). Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics*, 48:317–323.

Lehoczky, J. (1998). Markov chains. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics.* Wiley, Chichester.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, 5:81–91.

Lindgren, G. and Rychlik, I. (1991). Slepian models and regression approximations in crossing and extreme value theory. *International Statistical Review*, 59:195–225.

Luceño, A. (1993). A fast algorithm for the repeated evaluation of the likelihood of a general linear process for long series. *Journal of the American Statistical Association*, 88:229–236.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis.* Springer-Verlag, Berlin.

MacDonald, I. L. and Raubenheimer, D. (1995). Hidden Markov models and animal behaviour. *Biometrical Journal*, 37:701–711.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman & Hall, London.

McFadden, D. (1982). Qualitative response models. In Hildenbrand, W., editor, *Advances in Econometrics*, pages 1–37. Cambridge University Press, Cambridge.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36:318–324.

McNamara, J. M. and Houston, A. I. (1985). Optimal foraging and learning. *Journal of Theoretical Biology*, 117:231–249.

Montroll, E. W. and Shlesinger, M. F. (1982). On $1/f$ noise and other distributions with long tails. *Proceedings of the National Academy of Sciences of the USA - Applied Mathematical Sciences*, 79:3380–3383.

Nott, D. J. and Rydén (1999). Pairwise likelihood methods for inference in image models. *Biometrika*, 86:661–676.

Numerical Algorithms Group (1993). *Library Manual Mark 16.* Oxford.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, 57:99–138.

Phadke, M. S. and Kedem, G. (1978). Computation of the exact likelihood function of multivariate moving average models. *Biometrika*, 65:511–519.

Pyke, R. (1961). Markov renewal processes: definitions and preliminary properties. *Annals of Mathematical Statistics*, 32:1231–1242.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

Roberts, G. (1994). When to scan: an analysis of predictability in vigilance sequences using autoregression models. *Animal Behaviour*, 48:579–585.

Ross, G. J. S. (1998). Comparing the fit of non-nested non-linear models. In *Proceedings of Compstat98 conference.*

Sansom, J. (1999). Large-scale spatial variability of rainfall through hidden semi-Markov models of breakpoint data. *Journal of Geophysical Research*, 104(D24):31,631–31,643.

Sansom, J. and Thomson, P. J. (2000). Fitting hidden semi-Markov models. Technical Report 77, National Institute of Water and Atmospheric Research, Wellington, New Zealand.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Simpson, S. J. (1982). Patterns in feeding: a behavioural analysis using *locusta migratoria* nymphs. *Physiological Entomology*, pages 325–336.

Simpson, S. J. (1990). The pattern of feeding. In Chapman, R. F. and Joern, A., editors, *Biology of Grasshoppers.* John Wiley, New York.

Simpson, S. J. and Ludlow, A. R. (1986). Why locusts start to feed: a comparison of causal factors. *Animal Behaviour*, 34:480–496.

Slater, P. J. B. (1974). The temporal pattern of feeding in the zebra finch. *Animal Behaviour*, 22:506–515.

Slater, P. J. B. and Lester, N. P. (1982). Minimising errors in splitting behaviour into bouts. *Behaviour*, 79:153–161.

Slater, P. J. B. and Ollason, J. C. (1972). The temporal pattern of behaviour in isolated male zebra finches: transition analysis. *Behaviour*, 42:248–268.

Stuart, A., Ord, J. K., and Arnold, S. (1999). *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model.* Arnold, London, sixth edition.

Thuijsman, F., Peleg, B., Amitai, M., and Shmida, A. (1995). Automata, matching and foraging behavior of bees. *Journal of Theoretical Biology*, 175:305–316.

Titterington, D. M. (1990). Some recent research in the analysis of mixture distributions. *Statistics*, 21:619–641.

Tolkamp, B. J., Allcroft, D. J., Austin, E. J., Nielsen, B. L., and Kyriazakis, I. (1998a). Satiety splits feeding behaviour into bouts. *Journal of Theoretical Biology*, 194:235–250.

Tolkamp, B. J., Dewhurst, R. J., Friggens, N. C., Kyriazakis, I., Veerkamp, R. F., and Oldham, J. D. (1998b). Diet choice by dairy cows. 1. Selection of feed protein content during the first half of lactation. *Journal of Dairy Science*, 81:2657–2669.

Tolkamp, B. J. and Kyriazakis, I. (1997). Measuring diet selection in dairy cows: effect of training on selection of dietary protein level. *Animal Science*, 64:197–207.

Tolkamp, B. J. and Kyriazakis, I. (1999a). A comparison of five methods that estimate meal criteria for cattle. *Animal Science*, 69:501–514.

Tolkamp, B. J. and Kyriazakis, I. (1999b). To split behaviour into bouts, log-transform the intervals. *Animal Behaviour*, 57:807–817.

Tolkamp, B. J., Schweitzer, D. P. N., and Kyriazakis, I. (2000). The biologically relevant unit for the short-term feeding behavior of dairy cows. *Journal of Dairy Science*, 83:2057–2068.

Tong, Y. L. (1990). *The Multivariate Normal Distribution.* Springer, New York.

Victoria-Feser, M.-P. (1997). A robust test for non-nested hypotheses. *Journal of the Royal Statistical Society. Series B*, 59:715–727.

White, H. (1982). Regularity conditions for Cox's test of non-nested hypotheses. *Journal of Econometrics*, 19:301–318.

Whittle, P. (1953). The analysis of multiple stationary time series. *Journal of the Royal Statistical Society, Series B*, 15:125–139.

Williams, D. A. (1970). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, 28:23–32.

Yeates, M. P., Tolkamp, B. J., Allcroft, D. J., and Kyriazakis, I. (2001). The use of mixture distribution models to determine bout criteria. *Journal of Theoretical Biology*. (submitted).